

Supplementary materials for: The genome of the choanoflagellate *Monosiga brevicollis* and the origins of metazoan multicellularity

Nicole King^{1,2}, M. Jody Westbrook^{1*}, Susan L. Young^{1*}, Alan Kuo³, Monika Abedin¹, Jarrod Chapman¹, Stephen Fairclough¹, Uffe Hellsten³, Yoh Isogai¹, Ivica Letunic⁴, Michael Marr⁵, David Pincus⁶, Nicholas Putnam¹, Antonis Rokas⁷, Kevin J. Wright¹, Richard Zuzow¹, William Dirks¹, Matthew Good⁶, David Goodstein¹, Derek Lemons⁸, Wanqing Li⁹, Jessica Lyons¹, Andrea Morris¹⁰, Scott Nichols¹, Daniel J. Richter¹, Asaf Salamov³, JGI Sequencing³, Peer Bork⁴, Wendell A. Lim⁶, Gerard Manning¹¹, W. Todd Miller⁹, William McGinnis⁸, Harris Shapiro³, Robert Tjian¹, Igor V. Grigoriev³, Daniel Rokhsar^{1,3}

¹Department of Molecular and Cell Biology and the Center for Integrative Genomics, University of California, Berkeley, CA 94720, USA

²Department of Integrative Biology, University of California, Berkeley, CA 94720, USA

³Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA

⁴EMBL, Meyerhofstrasse 1, 69012 Heidelberg, Germany

⁵Department of Biology, Brandeis University, Waltham, MA 02454

⁶Department of Cellular and Molecular Pharmacology, University of California, San Francisco, San Francisco, CA 94158, USA

⁷Vanderbilt University, Department of Biological Sciences, Nashville, TN 37235, USA

⁸Division of Biological Sciences, University of California, San Diego La Jolla, CA 92093

⁹Department of Physiology and Biophysics, Stony Brook University, Stony Brook, NY 11794

¹⁰University of Michigan, Department of Cellular and Molecular Biology, Ann Arbor MI 48109

¹¹Razavi Newman Bioinformatics Center, Salk Institute for Biological Studies, La Jolla, CA 92037

*These authors contributed equally to this work.

Contents:

Supplementary Figures

- Figure S1: Choanoflagellates are a close outgroup of Metazoa
- Figure S2: Distribution of intron lengths in *M. brevicollis*
- Figure S3: Analysis of intron evolution in nine species
- Figure S4: Analysis of intron evolution in five species
- Figure S5: Domains significantly over-represented in choanoflagellates
- Figure S6. Legend for domains shown in Figure 4 - Domain shuffling and the evolution of Notch and Hedgehog.
- Figure S7: MbSrc functions like human c-Src
- Figure S8: Diagrams of metazoan general transcription factors and coactivators
- Figure S9: TBP-related factor in *M. brevicollis*
- Figure S10: Relative abundance of transcription factor families in *M. brevicollis*
- Figure S11: Alignment of homeodomain sequences used for Mr. Bayes analysis
- Figure S12: Phylogenetic relationships of human, sponge, choanoflagellate and fungal homeodomains

Supplementary Tables

- Table S1: Genome sequencing summary
- Table S2: Supporting evidence for gene models
- Table S3: Intron gain and loss as calculated by Csuros maximum likelihood
- Table S4: Functional classification of domains unique to choanoflagellates and metazoans
- Table S5: Protein domains unique to *M. brevicollis* and other groups
- Table S6: Species included in initial protein domain analysis
- Table S7: Immunoglobulin domains are restricted to choanoflagellates and animals
- Table S8: Intercellular signaling pathways across phyla
- Table S9: *M. brevicollis* presents a key intermediate in the evolution of MAPK signaling
- Table S10: Basal transcription factors present in *M. brevicollis*
- Table S11: Number of *M. brevicollis* protein models containing transcription factor family specific domains

Supplementary Notes

- S1. Genome sequencing and assembly
 - S1.1 Pilot sequencing efforts
 - S1.2 Generation of a monoxenic *M. brevicollis* culture, MX1
 - S1.3 Isolation of *M. brevicollis* genomic DNA
 - S1.4 Genome assembly and validation
 - S1.5 Assembly analysis and quality control
 - S1.6 No detectable single nucleotide polymorphism in *M. brevicollis*
 - S1.7 Mode of reproduction and ploidy of *M. brevicollis* remain unknown
- S2. Joint Genome Institute annotation of the genome

- S3. Analysis with an evolutionary perspective
 - S3.1 Phylogenetic Analysis
 - S3.2 Gene structure statistics
 - S3.3 Intron evolution
 - S3.4 Protein domain content of *M. brevicollis*
 - S3.5 Analysis of signaling, adhesion and transcription factor families
 - S3.6 Protein identification numbers for *M. brevicollis* and metazoan signaling homologs
 - S3.7 Phospho-tyrosine signaling
 - S3.8 TATA-binding proteins and transcription elongation factors
 - S3.9 MAPK signaling
- S4. Immunofluorescence Staining of *M. brevicollis*
- S5. Resources for choanoflagellate genomics

References for Supplementary Materials

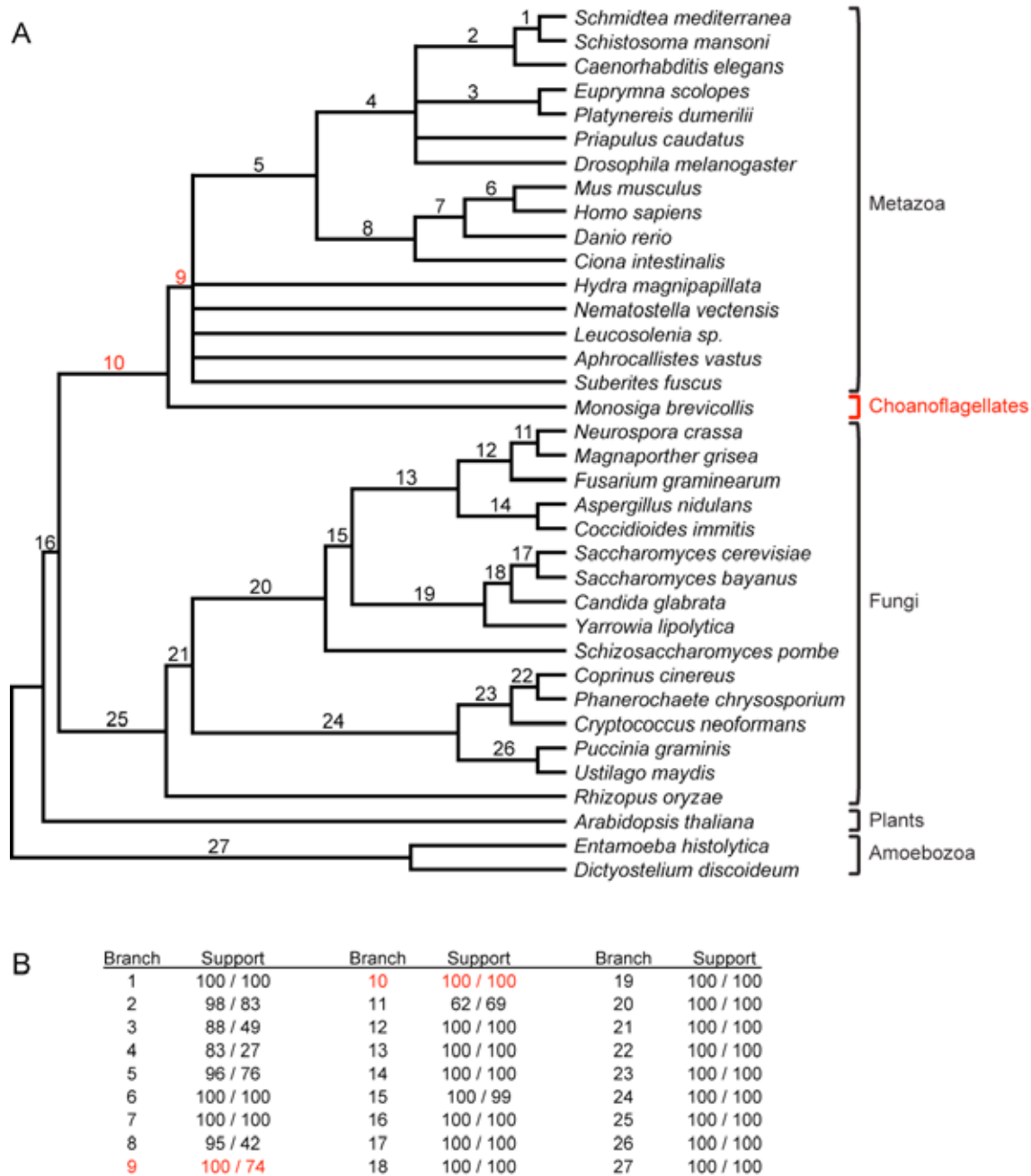
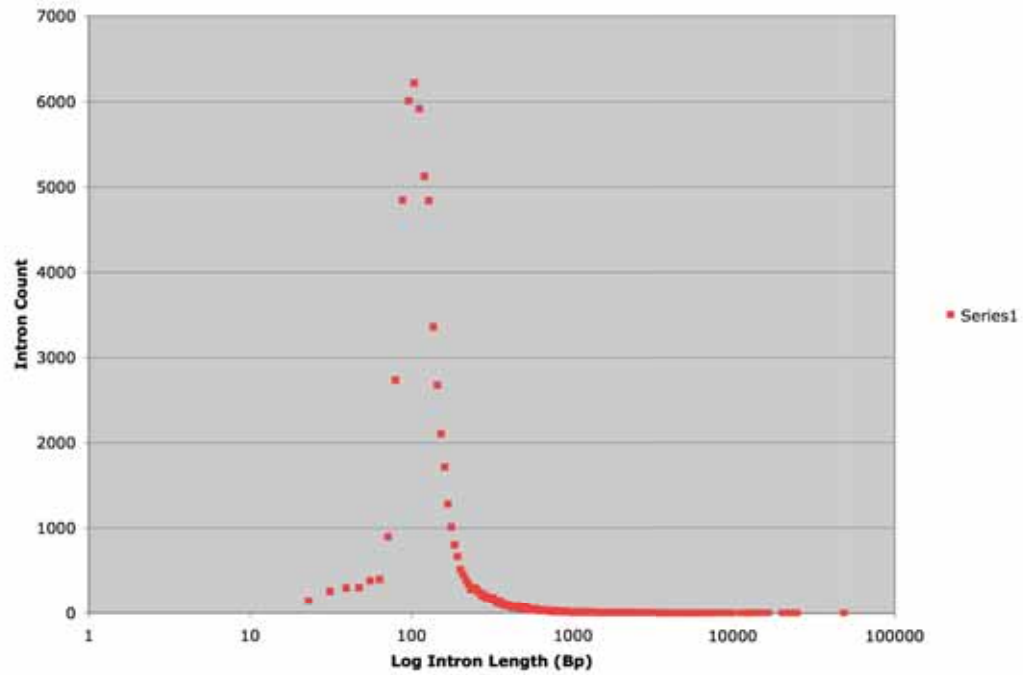


Figure S1. Choanoflagellates are a close outgroup of Metazoa. A phylogenetic analysis of 50 genes shows that *M. brevicollis* is placed outside metazoans (including poriferans and cnidarians), and justifies its choice for comparative genomic investigations into the transition from a unicellular to the multicellular metazoan lifestyle. (A) The tree with the highest likelihood in the maximum likelihood analyses is shown. (B) Bootstrap support values for all branches shown in A are shown. For each branch, the bootstrap support values from the maximum likelihood and maximum parsimony are shown, respectively.

A.



B.

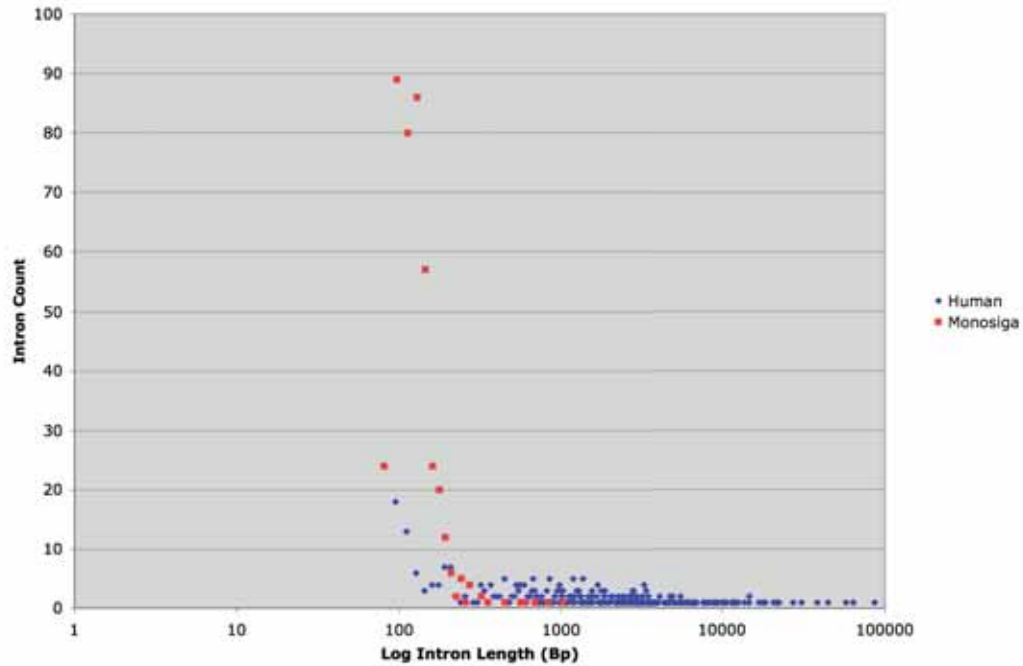


Figure S5. Distribution of *M. brevicollis* intron lengths. A. Distribution of the lengths of the 60,636 introns from the *M. brevicollis* filtered gene models. B. Distribution of the lengths of 419 introns that occur at the same positions in orthologous genes in *M. brevicollis* and humans.

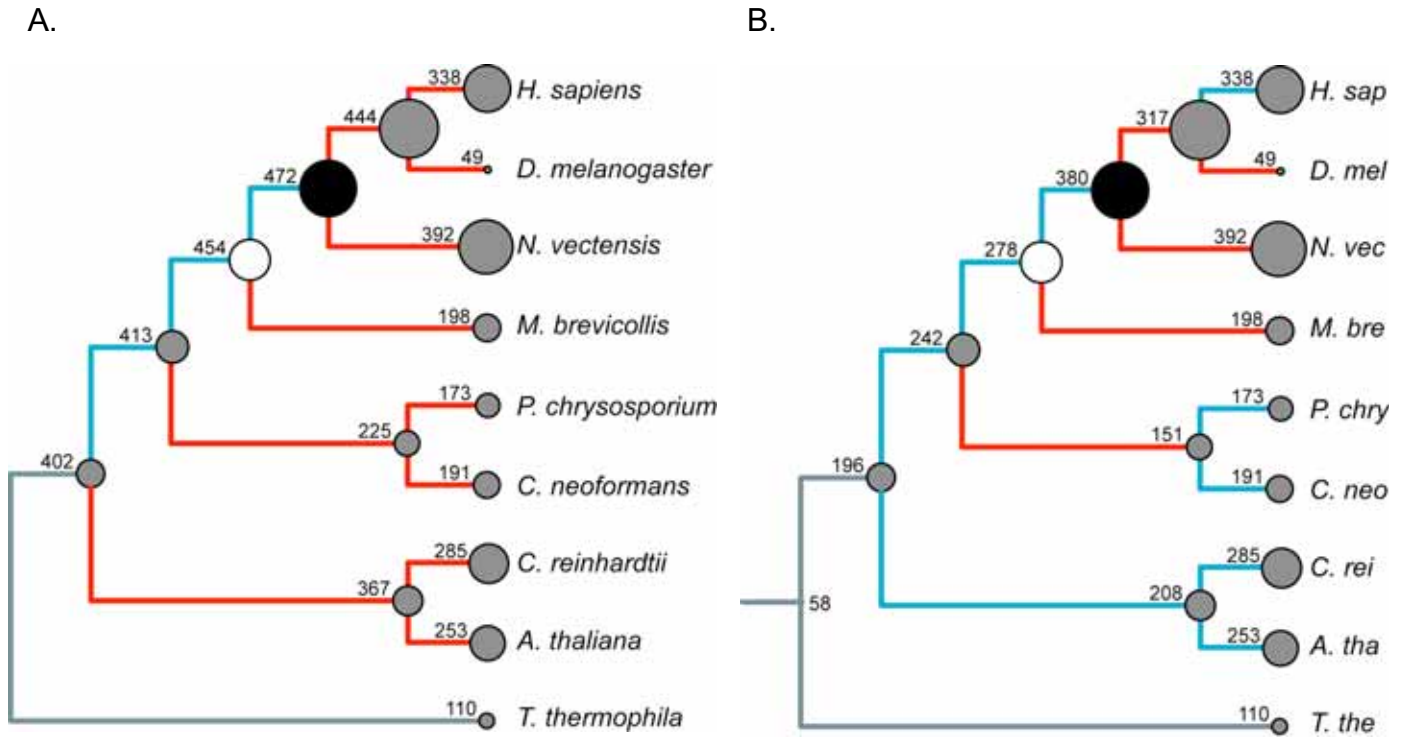


Figure S3. Analysis of intron evolution in nine species. Ancestral intron content and intron gains and losses were inferred using two additional methods: A. Roy-Gilbert maximum likelihood and B. Dollo parsimony methods. A sample of 1,054 intron positions in highly conserved sequences from 473 orthologs were used. Branches with at least 10% more gain than loss are blue, those with more loss than gain are red, and those with comparable amounts are black. Outgroup branches, for which intron loss could not be calculated, are grey. The inferred or observed number of introns present in ancestors and extant taxa are next to proportionally sized circles. Species included are *Tetrahymena thermophila* (*T. the*), *Chlamydomonas reinhardtii* (*C. rei*), *Arabidopsis thaliana* (*A. tha*), *Cryptococcus neoformans* A (*C. neo*), *Phanerochaete chrysosporium* (*P. chr*), *Monosiga brevicollis* (*M. bre*), *Nematostella vectensis* (*N. vec*), *Drosophila melanogaster* (*D. mel*) and humans (*H. sap*).

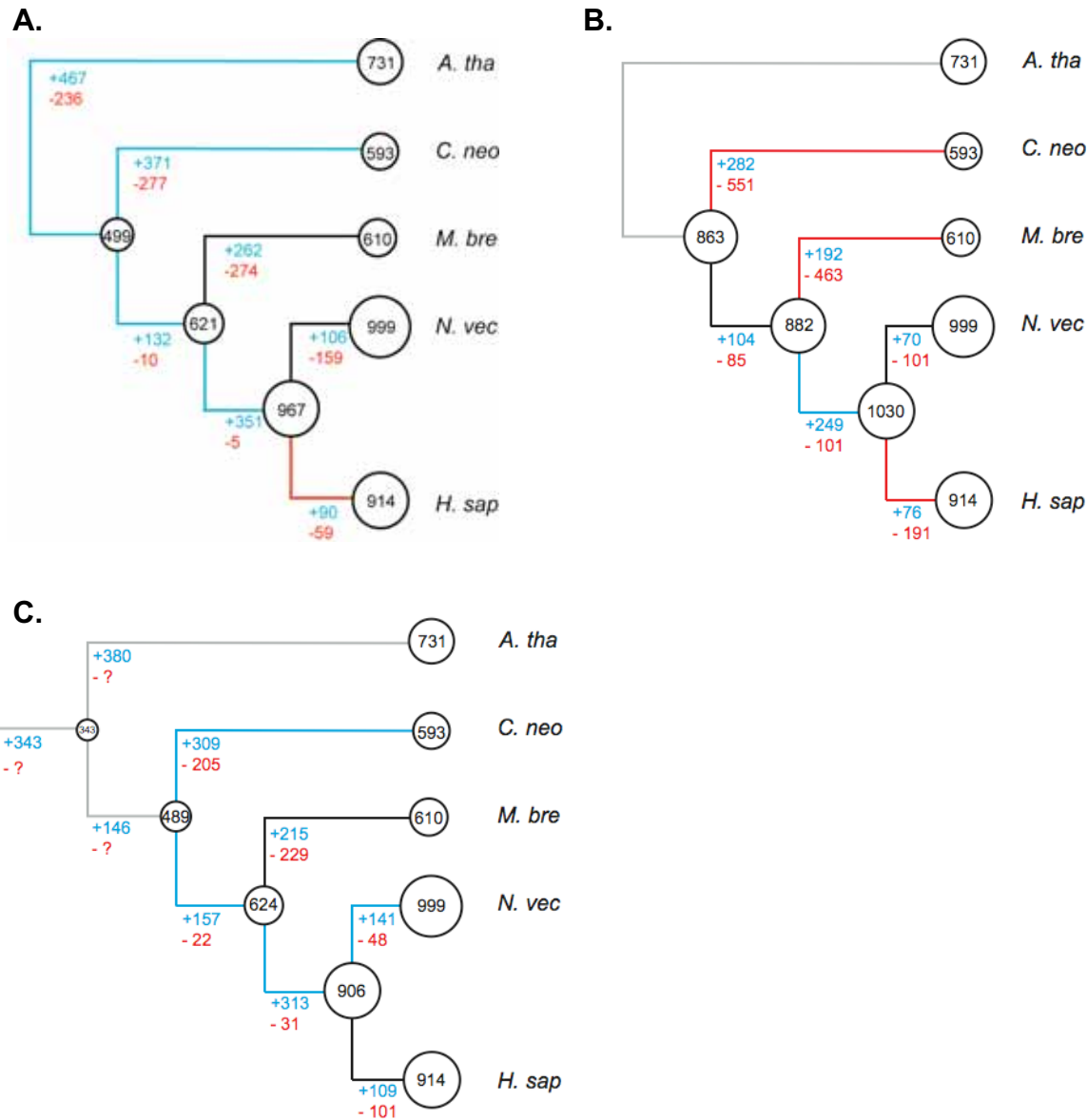
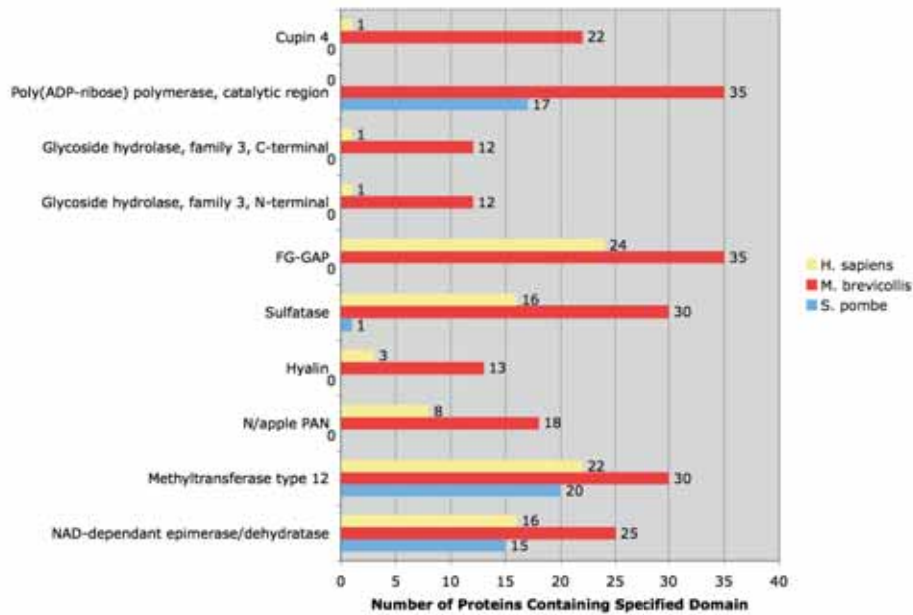


Figure S4. Analysis of intron evolution in five species. Ancestral intron content and intron gains and losses were inferred using three methods: **A.** Csuros maximum likelihood, **B.** Roy-Gilbert maximum likelihood and **C.** Dollo parsimony methods. A sample of 2121 intron positions in highly conserved sequences from 538 orthologs were used. Branches with 10% more gain than loss are blue, those with more loss than gain are red, and those with comparable amounts are black. Outgroup branches are grey. The numbers of introns gained and lost are shown in blue and red respectively. Using Dollo parsimony, the number of introns lost cannot be inferred without an outgroup, and this is indicated by question marks. The inferred or observed number of introns present in ancestors and extant taxa are in proportionally sized circles. Species included are the plant *Arabidopsis thaliana* (*A. tha*), the fungus *Cryptococcus neoformans* A (*C. neo*), the choanoflagellate *M. brevicollis* (*M. bre*) and the metazoans *Nematostella vectensis* (*N. vec*) and humans (*H. sap*).

A.



B.

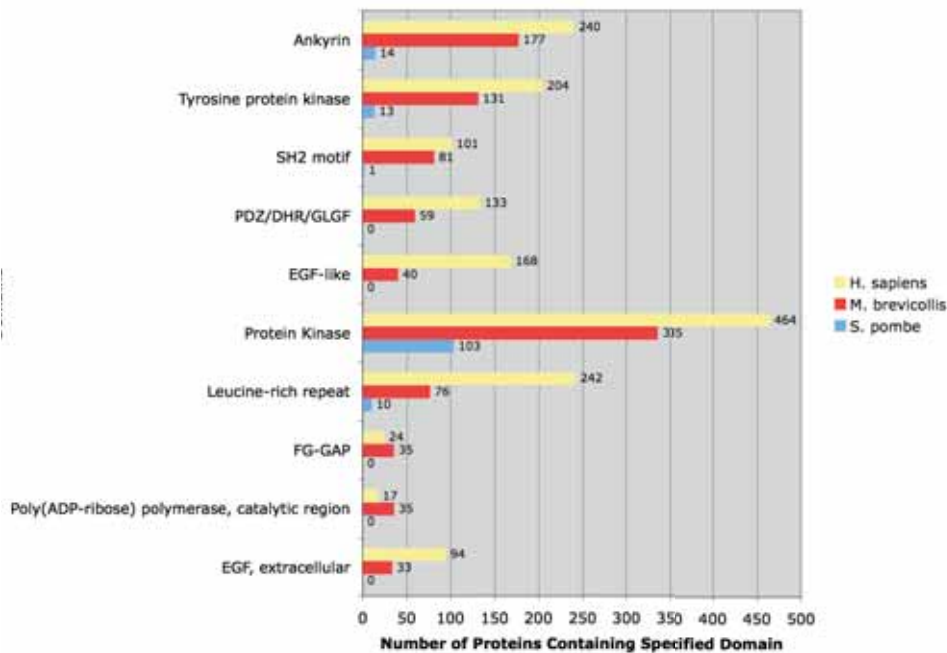


Figure S5. Domains significantly over-represented in choanoflagellates.

Significantly over-represented domains in the choanoflagellate genome were identified by comparing the occurrence of PFAM domains excluding repeats (one hit per protein) in *M. brevicollis* to the human (panel A) and *S. pombe* (panel B) genomes. The ten most significantly over represented domains from each comparison as determined by a Chi-squared test are shown, with the most significantly over-represented domain shown at the top of the graphs. The number of proteins containing each domain is indicated.



Figure S6. Legend for domains shown in Figure 4 - Domain shuffling and the evolution of Notch and Hedgehog. Analysis of the draft gene set reveals that *M. brevicollis* possesses protein domains characteristic of metazoan Notch and Hedgehog (Hh) proteins, some of which were previously thought to be unique to metazoans. The presence of these domains in disparate peptides in *M. brevicollis* suggests that domain shuffling has occurred in these proteins since the separation of the choanoflagellate and metazoan lineages.

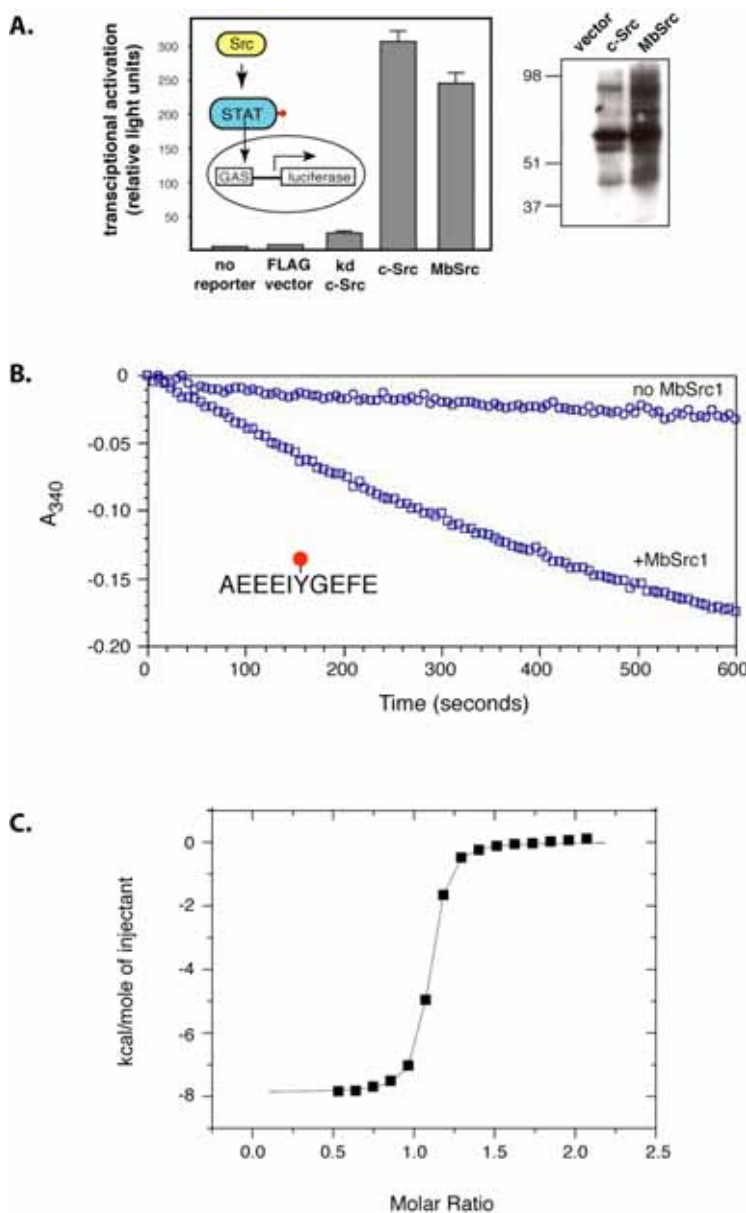


Figure S7. MbSrc functions like human c-Src. A. MbSrc can substitute for c-Src in a reporter assay. Src/Fyn/Yes triple knockout (SYF) cells were transfected with the indicated FLAG-constructs and with a luciferase reporter gene regulated by the interferon-gamma activation sequence. kd = kinase-dead c-Src. B. MbSrc phosphorylates substrates in mammalian cells. SYF cells were transfected with wild-type c-Src, Y527F c-Src, or MbSrc. Tyrosine-phosphorylated proteins in whole cell lysates were visualized by anti-pY Western blotting. C. Kinase activity of purified MbSrc. MbSrc was expressed and purified using the Sf9/baculovirus system. Phosphorylation of a synthetic peptide substrate containing the Src optimal motif was measured by a continuous spectrophotometric assay.

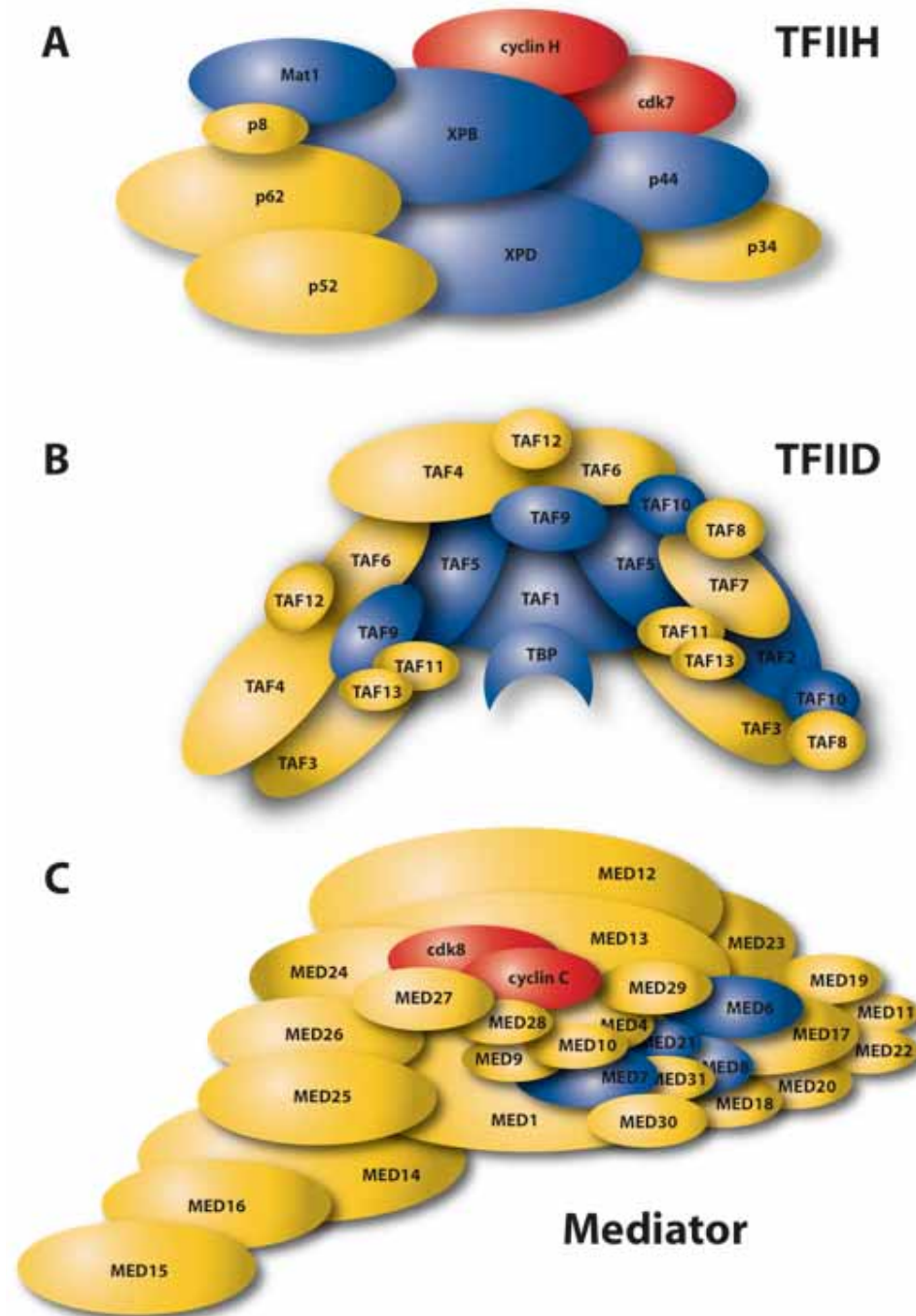


Figure S8. Diagrams of metazoan general transcription factors and coactivators. Blue indicates subunits found in *M. brevicollis*; yellow indicates a subunit not found in *M. brevicollis*; and red indicates a possible homolog in *M. brevicollis*. A. Diagram of TFIIF. B. Diagram of TFIID. C. Diagram of Mediator.

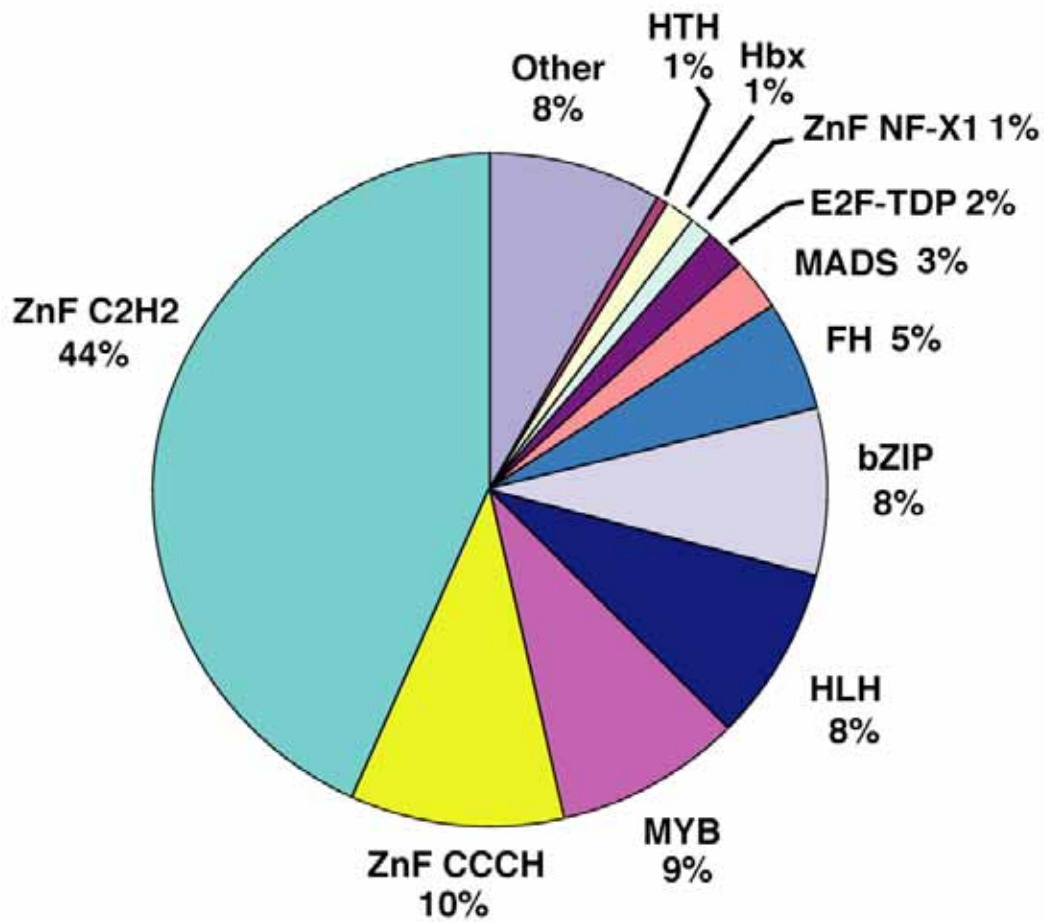


Figure S10. Relative abundance of transcription factor families in *M. brevicollis*. Of 155 protein models containing transcription factor associated domains, the percentage of protein models containing the indicated family specific domain is shown. bZip: basic-leucine zipper; E2f-TDP: E2F/DP (dimerization partner) family winged-helix DNA-binding domain; FH: forkhead; Hbx: homeobox; HLH: helix-loop-helix; HTH: helix-turn-helix; ZnF: zinc finger.

HESX_HUMAN GRRPRTAFTQNQIEVLENVF~~~RVNCYPGIDIREDLAQKLNLEEDRIQIWFQNRRRAKLKRSH
PMXA_HUMAN QRRIRTTFTSAQLKELELVF~~~AETHYPDIYTREELALKIDLTEARVQVWFQNRRRAKFRKQE
PMX1_HUMAN QRRNRTTFNSSQLQALERVF~~~ERTHYPDAFVREDLARRVNLTEARVQVWFQNRRRAKFRNE
OTX1_HUMAN QRRERTTFTSRQLDVLEALF~~~AKTRYPDIFMREEVALKINLPESRVQVWFKNRRRAKCRQQQ
CRT1_HUMAN KRRHRTTFTSLQLEELEKVF~~~QKTHYPDVYVREQLALRTELTEARVQVWFQNRRRAKWRKRE
PRH1_HUMAN RRRHRTTFTSPVQLEQLESAF~~~GRNQYPDIWARESLARDTGLSEARIQVWFQNRRRAKQRKQE
PIX1_HUMAN QRRQRTHFTSQQLQELEATF~~~QRNRYPDMSMREEIAVWNTLTPRVRVWFKNRRRAKWRKRE
GSC_HUMAN KRRHRTIFTDEQLEALENLF~~~QETKYPDVGTREQLARKVHLREEKEVWFKNRRRAKWRKQK
PAX6_HUMAN LQRNRTSFTQEIEALEKEF~~~ERTHYPDVFAERERLAAKIDLPEARIQVWFNSNRRRAKWRREE
Renprd1 QRRHRTNFTSHQLEELEKAF~~~EKTRYPDVFMREELAMKISLTEARVQVWFQNRRRAKWRKAE
Renprd2 SKRNRTTFTAHQLDELEMIF~~~RQTHYPDVLLREKLAQRIGLPESRVQVWFQNRRRAKWRKRE
Renprd3 KRRYRTTFTSFQLELELEKVF~~~ERTHYPDVFTREDLANRVELTEARVQVWFQNRRRAKWRKKE
Renprd4 QRRFRRTTFTSYQLQELEAAF~~~AKTHYPDVFMREDLALRINLTEARVQVWFQNRRRAKWRRAQ
Renprd5 PKRTRTAYSNSQLDQLELIF~~~ATTHYPDVFTREDLSRRLGIREDRIQVWFQNRRRAKFRKQE
Renprd6 IKKKRMTYTKQKDALESYF~~~YQDSYPTQARENMSEALGITPEKVQVWFQNRRRAKCRKRE
Renprd7 PKKTRTQFSPKQLVYLEECF~~~LKNRFPsAKERESIAEELDLTQHIIQVWFQNRRRAKHRRKS
LHX2_HUMAN TKRMRTSFKHHQLRMTKSYF~~~AINHNPDADKDLKQLAQKTGLTKRVLQVWFQNARAKFRNL
LH61_HUMAN AKRARTSFTAQQLQVMAQF~~~AQDNNDPAQTLQKLADMTGLSRRVQVWFQNCRARHKHT
ISL1_HUMAN TTRVRTVLNEKQLHTLTCTY~~~AANRPDALMKEQLVEMTGLSPRVIRVWFQNRRCKDKKRS
LHX3_HUMAN AKRPRTTITAKQLETLKSAY~~~NTSPKPARHVREQLSSETGLDMRVVQVWFQNRRRAKEKRLK
LMXB_HUMAN PKRPRTILTQQRRAFKASF~~~EVSSKPCRKVRETLAAETGLSVRVVQVWFQNQRAKMKKLA
RenLIM1 KGKTRTSINPKQLIVLQATY~~~EKEPRPSRSMREELAAQTGLTAKVIQVWFQNRRSKDKKDG
RenLIM2 QPRIRTVLTEQQLQTLRSVY~~~QTNPRPDALLKEQLCELTGLSPRVIRVWFQNRRCKDKKAL
RenLIM3 QKRPRTTISQKQLDLLKTAY~~~CVSPKPSRHVRQELSDKTGLDMRVVQVWFQNKRAKDKRTK
OCT6_HUMAN KRKKRTSIEVGKGALESHF~~~LKCPKPSAHEITGLADSLQLEKEVVRVWFQCNRRQKEKMT
PO61_HUMAN KRKKRTSFTPAIEALNAYF~~~EKNPLPTGQEITEIAKELNYDREVRVWFQCNRRQTLKNTS
BR3A_HUMAN KRKKRTSIAAPEKRSLEAYF~~~AVQPRPSSEKIAAIAEKLDLKKNVVRVWFQCNRRQKQKRMK
OC3A_HUMAN KRKKRTSIENRVRGLENLF~~~LQCPKPTLQQISHIAQQLGLEKDVVRVWFQCNRRQKGRSS
RenPOU1 HRKKRTTIGMSAKERLEQHF~~~VQVQPKSSSDITKVADSLNLDKEVIRVWFQCNRRQREKVR
RenPOU2 RRRRRTAIPVQTKKQLLKEF~~~ENNPKPSVKALKALAEKLGIRFEVVRVWFQCNRAKKKAGK
RenPOU3 KRKGRTAISVQTKKQLLKEF~~~ENDPKPSPKDLKAISEKLIGFEVVRVWFQCNRAKKKAGK
RenPOU4 KRKKRVVYTPHALSILNKYF~~~LKEPRPNRQIEMVAEELDLLPEEVVRVWFQCNRRQKYEKMT
A.nid1 KNNKRQRATQDQLVLEMEF~~~NKNPTPTAATRERIAQEINMTERSVQIWFQNRRRAKIKMLA
N.cra1 KNQKRQRATQDQLTLEMEF~~~NKNPTPTATVRERIAEEINMTERSVQIWFQNRRRAKIKLLA
R.ory1 STRKRTHLSTEQVSLLESSF~~~NENSLPDSAVRSRLAQELSVTERTVQIWFQNRRRAKEKKIK
P.bla3 AKPKRKRI SPDQFRVLSDLF~~~EKTDTPNYELRERMAGRLNMTNREVQVWFQNRRRAKATRAK
R.ory8 IRPKRKRI TPNQLEVLTSIF~~~ERTKTPNYQLREHTAKELNMTNREVQVWFQNRRRAKLNKR
R.ory2 RTRKRTRATPEQLALEKSF~~~NVNPSNSRVREQLSLQLGMTERSIIQIWFQNRRRAKVNQT
P.bla1 QPRKRTRASPEQLGILEKTF~~~NINPSNNRVREQLSQQLSMSERSIIQIWFQNRRRAKVNKIA
R.ory3 PVRKRTRATADQLSVLEDTF~~~AMNVSPNSKLRQLAEQLQMSERSIIQIWFQNRRRAKVKHM
R.ory4 DTKKRTRVTPGQLALEETF~~~SMTATPDSKLRQLAERLKMERSIIQIWFQNRRRAKVKMLQ
L.bic3 EKRRKRSVTEQQLVHLEQYF~~~KADRCPTATRRREISEQLGMQERQTQIWFQNRRRAKAKLQ
P.chr3 EQKKRGRVTPEQLAVLEAIF~~~AANRSPNAVRRKEISEQLGMTERQTQIWFQNRRRAKEKHAG
R.ory5 EIKHRRRTSRAQLKVLEESF~~~SENPKPNATVRRILAQQLDMTPRGVQIWFQNRRRAKALLR
R.ory6 ETKHRRRTSRGQVKILEKAF~~~HDNPKPNGRARERLAESLSMSPRGVQIWFQNRRRAKAKNQ
L.bic1 EVKHKRRTTSAQLKVLETVF~~~KRDTPKNASLRTELAAQLDMTARGVQVWFQNRRRAKEKVKA
R.ory7 IKAKRKRASPSQLYILNQVF~~~QQTCTPSTELRIELGKRLGMSPTVQIWFQNKRQSTRTKE
A.nid3 ARQKRRRTSPEDYAILEAEY~~~QRNPKPKDISRASIVSRVSLGEKEVQIWFQNRRQNDRRKS
N.cra3 PKGKRKRRTTAKDKAILEAAY~~~NANPKPKDKAARQDIVNRVSLNEKEVQIWFQNRRQNDRRKS
A.nid2 ENLSRPRLTKEQVETLEAQF~~~QAHPKPSSNVKRQLAQQLTHLSLPRVANWFQNRRRAKAKQKQ
N.cra2 QTEPKPRLAKDEVELLEREF~~~AKNPKPNTSLKRELAEQMGVEVPRINNWFQNRRRAKEKQMR
P.bla2 FHKKRMMLKPYQYKVLQDHF~~~SANPKPDARVYIDIASRLNVSITKIKNWFQNRRRAKARKDK
P.bla4 KIKNRRRFSATEAALLERRY~~~AEEQSPSQHVQLGLADQMSTPRKTTITWFQNRRRAKYKRS
P.bla5 EIKHRRHFSTSELELLEELY~~~RRHPRPSSEKKAMAALDTPGRVQVWLQNRRRAKERKAQ
R.ory9 PIKQRRRFSLEEAQFLEMEY~~~NNNPSPTQDKIQQIASKINSRKVVTTWFQNRRRAKNRRRS
R.ory10 PIRPRKRFTSNQIHLEMEY~~~MKS DHPSRETKETLANQFKTSIRRIQIWFQNRRRAKEKRGE
R.ory11 VARRRMRTSKEEMAVLDEYY~~~RKNPNPNQEEKKEIANLLKMGTKNVHFWFQNRRRAKKNKK
A.nig1 KMKRFRRLTHNQTRFLMSEF~~~TRQAHPDAAHRERLSKEIGLTPRQVQVWFQNRRRAKLRLT
N.cra4 RKMKFRRLTHQTRFLMSEF~~~AKQPHPDAAHRERLSREIGLSPRQVQVWFQNRRRAKIKRLT
C.neo3 QVKHRRRTTPEQLKVLEFWY~~~DINPKPDNLREQLAAQLGMTKRNQVWFQNRRRAKMKGLA
C.neo4 FKSPRKRTNDVQLAMLSEVF~~~RRTQYPSTEERDELAKQLGMTSRVQIWFQNRRRAVKVDQ

Figure continued on next page

P.chr2	EKKPRHRMTDKQLERLEALY~~~QQDTHPTREQKQALGEEVGMDTRTVTVWFQNRRLQLSKKNT
C.neo1	KMSPRKRFTIPQLQILEVQW~~~SNDISPPKVDRQLAMWMGTRTKHVNIWFQNRRLQYEKKVH
C.neo2	GCKVRRRFTKRELEALEVLW~~~SIAKSPSKYERQRLGAWLGVKTKHITVWFQNRRLQEEKRY
L.bic2	IRKKRKRVDAAQLKVLNETY~~~NRTAFPPSTEERHTLAKALDMSARGVQIWFQNRRLQSRQTN
C.cin1	SRRTKRFTNTQLTMLNLNF~~~HQTSHPSREEREAVAKAGQMEIKSVTIWFQNRRLQTERKSQ
P.chr1	PKKPRHRHSAFQLAALNELY~~~ERDEHPPLEERTSLAERLGMEVKTVNAWFQNRRLQSTKKRS
P.chr4	VSYGRRRMQPEQLQALQTLY~~~DANTHPTKAQRMQLARELDLDLKSVDNVWYQNRRLQSMKKKL
P.bla6	IAKRRPRTTPEQSRILNTHF~~~ARNPVPSKNEIKLIAREVKIKPRSTHFWYQNRRLQSVKREG
CUT1_HUMAN	LKKPRVVLAPEEKEALKRAY~~~QQKPYPSPKTIEDLATQLNLKTSTVINWFHNYRSRIRREL
SIX1_HUMAN	GEETSYCFKEKSRGVLREWY~~~AHNPYPSPREKRELAEATGLTTTQVSNWFKNRRLQDRDRAAE
SIX3_HUMAN	GEQKTHCFKERTSRLLREWY~~~IQDPYPNPSKKRELAQATGLTPTQVGNWFKNRRLQDRDRAAA
RenSIX	GEETSYCFKEKSRVVLQWY~~~TKNAYPSPREKQLAEQTGLTTTQVSNWFKNRRLQDRDRAAE
PBX1_HUMAN	ARRKRRNFNKQATEILNEYFYSHLSNPYPSEEAKEELAKKCGITVSQVSNWFGNKRIRYKKN
RenPBX	ITRTRPVLTRNSLKVLEEWYECHELDHPYPTASQVEWLAQVSSLNTEQVKKWFGNKRSRKNTR
IRX2_HUMAN	DPAYRKNA TRDATA TLKAWLNEHRKNPYPTKGEKIMLAIIITKMTLTQVSTWTFANARRRLKKEN
RenIRO1	SAAGSITRRMRNTAVLVKWIEDHQSNPYPTKAEQYLAYYSGMNM TQLSTWTFANARRRIKKIG
RenIRO2	VQLASSRRRRRDATHLIEWLDLHQGNPYPTRVEKEQLVVISGMNFKQLNDWTFANARRNIRKVG
RenIRX3	EKGSSSPGSRWNTDVLALWITEHLQLPYPGKVEKQYLCFYSNMSMKQVSTYFANARR~~~~~
RenIRO4	CSNDMEARGSEGYKTSGEVVGHAHQTNPYPTKAEKECLAECGMSVKQLCTWFSNRRQIRKLG
RenIRO5	YDSPRYKLTPERAIPLIKWFEEHDKHPYPSRHEKMLLCQSTQTLTFTQVSTWTFANARRRMKK~~
TGIF_HUMAN	KRRRRGNLPKESVQILRDWLYEHRYNAYPSEQEKALLSQQTHLSTLQVCNWFINARRLLPDM
MEI1_HUMAN	RHKKRGIFPKVATNIMRAWLFQHLTHPYPSEEQKKQLAQDTGLTILQVNNWFINARRRIVQPM
RenMEIS	TGKKREKTSPASQKLLKEWLFHSRCPYPTEDDKQNLCRM TGLSLQQLNNWFINARRRILPQK
MONOSIGA_MEIS1	SRHCTKRFASSSIDTLKEWLFHAHTDRPYPTDQDKTELMQQTGLDLMQINNWFINARRLLVKV
MONOSIGA_MEIS2	NTGGRNNMPHEVTSRLKEWFFAHTSHPYPSEQKKRELASQCDLTLLQINNWFINARRRLNRP
A.nid4	NRRRRGNLPKPVTEILKAWFHAHLDPYPSEEDKQMLMSRTGLTINQISNWFINARRRHLPAL
N.cra5	KNKRRGNLPKEVTEKLYAWLYGHLNHPYPTEDQKMMRETNMQMNQISNWFINARRRKVPLL
P.bla7	KKRRRGNLPREVTEFLKHWLIQHKAHYPYSEKEKGDLACRTGLTVNQISNWFINARRRILQPM
L.bic4	PQRKRGLPKETTDYDLKAWLHRHSDHPYPSEDEKKQLCHATGLSMSQVSNWMINARRRILAPA
N.cra6	ATKVNNRFSRESIKILKNWLSIHQKHPYPNDEEKEMLQKQTGLSKTQITGWLANARRRRGKVM
A.nid5	ARKSSSRLSREAVRILKAWLNDHSDHPYPTEEEKEELKLRTGLKRTQITNWLANARRRGKIRP
A.nid6	DSKESKQFVRKGARVL RDWIFYQNEHCPYPSEEEKARLAAETGFSRQRISTWTFANARRRHKQQK

Figure S11. Alignment of homeodomain sequences used for Mr. Bayes analysis. *Homo sapiens* homeodomain sequences were taken from the NCBI homeodomain resource. Sponge sequences are labeled with Ren and were found by BLAST of the *Reniera sp.* trace data from the NCBI trace archives. Fungal sequences were obtained from the Broad Institute (A.nid - *Aspergillus nidulans*; C.cin - *Coprinus cinerea*; C.neo - *Cryptococcus neoformans*; N.cra - *Neurospora crassa*; R.ory - *Rhizopus oryzae*) and JGI (A.nig - *Aspergillus niger*; L.bic - *Laccaria bicolor*; P.chr - *Phanerochaete chrysosporium*; P.bla - *Phycomyces blakesleeanus*).

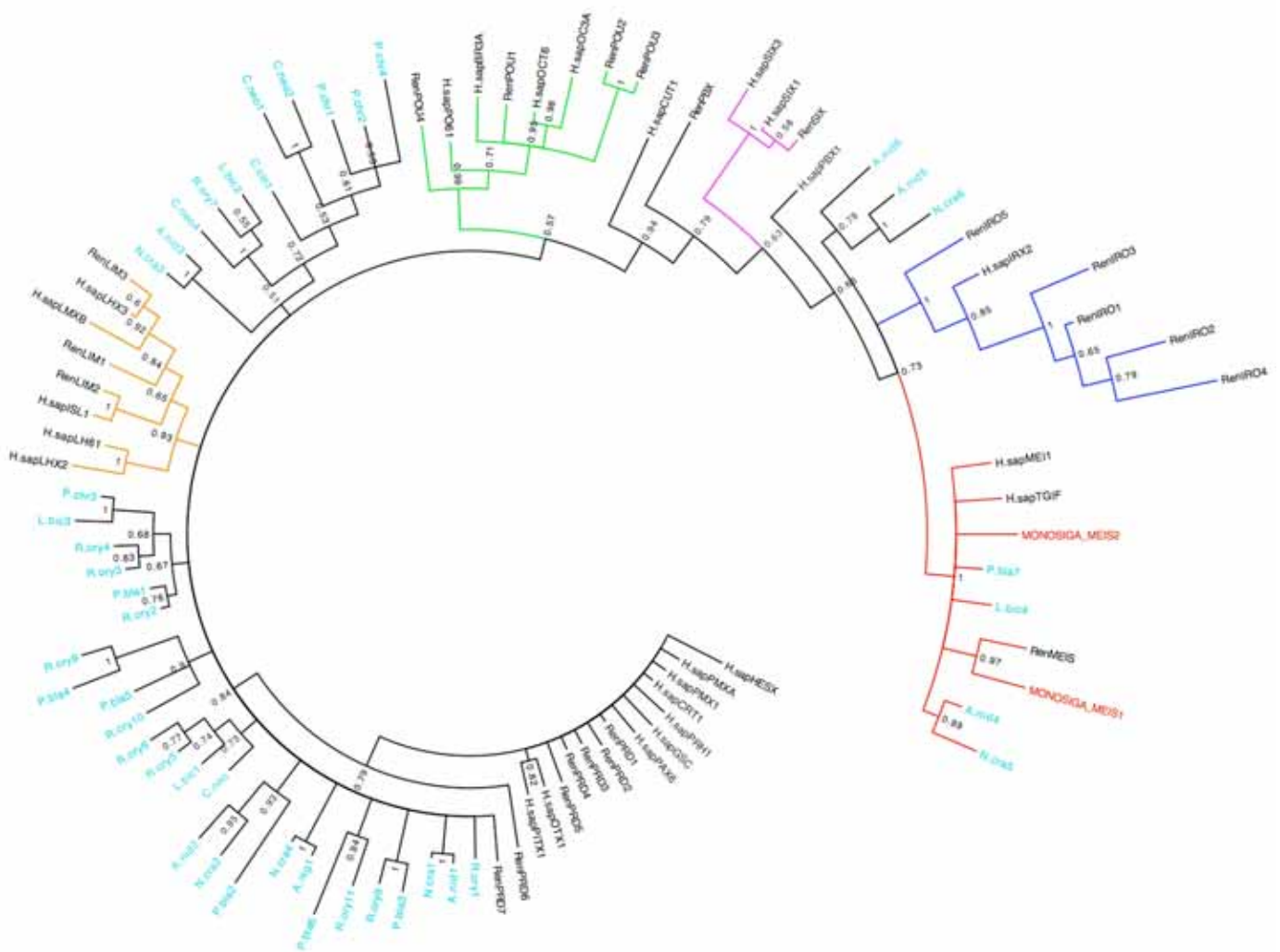


Figure S12. Phylogenetic relationships of representative human, sponge, and fungal homeodomains with the two *M. brevicollis* homeodomains.

Analysis was done with Mr. Bayes^{2,3} run with mixed amino acid and inverse gamma settings for 3 million iterations with a burnin of 75,000. The Tree was made using FigTree (Andrew Rambaut, <http://tree.bio.ed.ac.uk/>). Fungal gene labels are in light blue and those from *M. brevicollis* are labeled in red. MEIS class clade is highlighted in red, IRO in dark blue, SIX in purple, POU in green, and LIM in orange.

Table S1. Genome sequencing summary.

Library IDs	Theoretical insert size	Actual insert size	Raw reads	Raw (untrimmed) sequence (Mb)	Passing reads	Quality and vector trimmed sequence (Mb)
AZSO	2-3 kb	3,061 +/- 525	7,620	8	6,599	5
BHUH	2-3 kb	2,365 +/- 355	295,882	314	262,757	185
BAFY	6-8 kb	6,593 +/- 1,284	7,680	8	5,457	4
BNUS	6-8 kb	7,059 +/- 1,769	242,175	235	226,029	165
BAFZ	35-40 kb	38,665 +/- 11,944	3,840	4	3,308	2
BIFH	35-40 kb	36,888 +/- 13,666	77,856	76	46,940	22
Total			635,053	645	551,090	383

Table S2. Supporting evidence for genes models.

Evidence	<i>M. brevicollis</i> v.1
Complete models (annotated start and stop codons)	8286 (90%)
Models with EST alignment	4186 (46%)
Models with nr alignment (e-value < 0.1)	7590 (83%)
Models with Swissprot alignment (e-value < 10 ⁻⁵)	5877 (64%)
Models with Pfam alignment (gathering threshold)	5160 (56%)

Table S5. Intron gain and loss as calculated by Csuros maximum likelihood.

Branch	Introns Gained	Introns Lost
Eukaryotic → <i>T. the</i>	64	157
Eukaryotic → Green plants ancestor	65	52
Green plants ancestor → <i>A. tha</i>	73	36
Green plants ancestor → <i>C. rei</i>	177	108
Eukaryotic → Opisthokont ancestor	56	23
Opisthokont → Basidiomycete ancestor	75	126
Basidiomycete ancestor → <i>C. neo</i>	87	80
Basidiomycete ancestor → <i>P. chr</i>	32	42
Opisthokont → Holozoan ancestor	61	0
Holozoan ancestor → <i>M. bre</i>	69	167
Holozoan → Eumetazoan ancestor	135	23
Eumetazoan ancestor → <i>N. vec</i>	12	29
Eumetazoan → Bilaterian ancestor	30	13
Bilaterian ancestor → <i>D. mel</i>	21	397
Bilaterian ancestor → <i>H. sap</i>	1	89

Branches shown on the tree in Figure 2 are indicated by the ancestor or extant species at the end of the branch and the ancestor at the last bifurcation. Intron gains and losses were calculated by the Csuros intronRates program⁴ with no missing sites assumed and using an unrooted species tree. Holozoan ancestor denotes the ancestor of choanoflagellates and animals. Opisthokont ancestor denotes the ancestor of fungi and holozoans.

Table S4. Functional classification of domains unique to choanoflagellates and metazoans.

Cell Adhesion and Extracellular Matrix	
Cadherin*	Laminin G*
CUB	Laminin N-terminal
Ependymin	Reeler
Fibrillar collagen C-terminal	Somatomedin B
HYR*	Von Willebrand D*
Kunitz/bovine pancreatic trypsin inhibitor*	
Signal Transduction	
Antistatin family	Nine cysteines of family 3 GPCR
BTK motif	Pacifastin inhibitor (LCMII)
C1q*	Phosphotyrosine binding (IRS-1 type)
CBL proto-oncogene N-term, domain 1	Phosphotyrosine interaction (PTB/PID)
CBL proto-oncogene N-term, EF hand-like	PI3-kinase family, p85-binding
CBL proto-oncogene N-term, SH2-like	Plexin
ECSIT	Raf-like ras-binding
Flotilin family	Renin receptor-like protein
GoLoco motif	S-100/ICaBP type calcium binding
Heme NO binding associated	Seven transmembrane receptor, secretin family
Hormone receptor	SH3 domain-binding protein 5 (SH3BP5)
L27	Spin/Ssty family
Low-density lipoprotein receptor class A	TNF (Tumor Necrosis Factor)
Cell Adhesion and Signal Transduction	
Leucine rich repeat N-terminal	Immunoglobulin I-set*
Immunoglobulin	Immunoglobulin V-set*
Immunoglobulin c-2*	
Transcriptional Control	
Mbt repeat	STAT protein, DNA binding
p53 DNA-binding **	Zinc finger, C2HC type
PET	
Cytoskeletal Associated	
Nebulin repeat	Repeat in HS1/cortactin
Filament	Sarcoglycan complex subunit protein
Transporters/Channels	
Dihydropyridine sensitive L-type calcium channel	Organic anion transporter polypeptide (OATP)
Inward rectifier potassium channel	Progressive ankylosis protein (ANKH)
Enzymes	
Aspartyl/asparaginyl beta-hydroxylase	Galactosyl transferase
DNaseIc*	Glycosyl hydrolase family 59*
Cu ₂ monooxygenase	Heparan sulfate 2-O-sulfotransferase*
Fzo-like conserved region	N-acetylglucosaminyltransferase-IV conserved reg.
Galactose-3-O-sulfotransferase	Phosphomevalonate kinase
Unknown	
Assoc. with transcription factors and helicases	PHR
Domain of unknown function (DUF758)	Protein of unknown function (DUF1241)
Domain of unknown function (DUF837)	Selenoprotein S (SeIS)
Fukutin-related	Translocon-associated protein, δ subunit precursor
Hormone-sensitive lipase (HSL) N-terminus	Tropomyosin
MOFRL family*	Uncharacterized protein family (UPF0121)
N-terminal domain in <i>C. elegans</i> NRF-6	

* Present in bacteria

** Partial domain present in *Zea mays* (Qi, 2003)

Table S5. Protein domains unique to choanoflagellates and other groups.

<i>Domain Name</i>	<i>Interpro ID</i>
Metazoa, Choanoflagellates, Fungi, and Dictyostelium	
Growth-Arrest-Specific Protein 2 Domain	IPR003108
Protein of unknown function (DUF1183)	IPR009567
Protein of unknown function (DUF1613)	IPR011671
Mss4 protein	IPR007515
UcrQ family	IPR004205
Diaphanous FH3 Domain	IPR010472
WSC domain	IPR002889
TAP C-terminal domain*	IPR005637
RasGAP C-terminus	IPR000593
GGL domain	IPR001770
Ras association (RalGDS/AF-6) domain	IPR000159
I/LWEQ domain	IPR002558
BTG family	IPR002087
Cysteine dioxygenase type I*	IPR010300
Fic protein family*	IPR003812
Fes/CIP4 homology domain (FCH)	IPR001060
GTPase-activator protein for Ras-like GTPase (Ras GAP)	IPR008936
RasGEF	IPR001895
RasGEF, N-terminal motif	IPR000651
Wiskott Aldrich syndrom homology region 2*	IPR003124
Alpha adaptin AP2, C-terminal domain	IPR003164
G-protein gamma like domain (GGL)	IPR001770
BTG domain	IPR002087
Metazoa, Choanoflagellates, and Fungi	
Arfaptin	IPR010504
ATP synthase D chain, mitochondrial (ATP5H)	IPR008689
Cation-dependent mannose-6-phosphate receptor	IPR000296
CP2 transcription factor family	IPR007604
CybS	IPR007992
Cytochrome c oxidase subunit Va	IPR003204
D-ala D-ala ligase C-terminus	IPR011095
Disintegrin	IPR001762
Dolichyl-phosphate-mannose-protein mannosyltransferase	IPR003342
Epoxide hydrolase N terminus	IPR010497
Forkhead domain	IPR001766
FRG1-like family	IPR010414
GDP/GTP exchange factor Sec2p	IPR009449
Golgi phosphoprotein 3 (GPP34)	IPR008628
HRDC (Helicase and RNase D C-terminal) domain	IPR002121
Inhibitor of Apoptosis domain	IPR001370
Microtubule associated	IPR012943
Peptidase C1-like family	IPR004134
Protein of unknown function (DUF1349)	IPR009784
Putative phosphatase regulatory subunit	IPR005036
Receptor L domain	IPR000494
RFX DNA-binding domain	IPR003150
SURF4 family	IPR002995
TEA/ATTS domain family	IPR000818
XPA protein C-terminus	IPR000465

XPA protein N-terminal	IPR000465
------------------------	-----------

Metazoa, Choanoflagellates, and Dictyostelium

Tryptophan 2,3-dioxygenase*	IPR004981
DUF1632	IPR012435
Beta catenin interacting protein (ICAT)	IPR009428
DUF1394	IPR009828
RUN domain	IPR004012
Doublecortin	IPR003533
Translocon assoc. protein, gamma subunit	IPR009779
Hyaluronidase 2*	IPR013618
DUF1736	IPR013618
Fascin*	IPR010431
IRSp53/MIM homology domain (IMD)	IPR013606
Survival motor neuron protein (SMN)	IPR010304
Spectrin	IPR002017
Translocon-assoc protein, gamma subunit (TRAP-gamma)	IPR009779
Follistatin-N-terminal domain-like (FOLN)*	IPR003645

Metazoa and Choanoflagellates

Antistasin family	IPR004094
Aspartyl/asparaginyl beta-hydroxylase	IPR007803
Associated with TFs and helicases	IPR006576
BTK motif	IPR001562
C1q*	IPR001073
Cadherin*	IPR002126
CBL proto-oncogene N-term, domain 1	IPR003153
CBL proto-oncogene N-term, EF hand-like	IPR003153
CBL proto-oncogene N-term, SH2-like	IPR003153
Collagen triple helix	IPR000087
Cu ₂ monooxygenase	IPR003153
CUB	IPR000859
Dihydropyridine sensitive L-type calcium channel	IPR000584
DNaseIc*	IPR008185
Domain of unknown function (DUF758)	IPR008477
Domain of unknown function (DUF837)	IPR008555
ECSIT	IPR010418
Ependymin	IPR001299
Fibrillar collagen C-terminal	IPR000885
Filament	IPR001664
Flotillin*	IPR004851
Fukutin-related	IPR009644
Fzo-like conserved region	IPR006884
Galactose-3-O-sulfotransferase	IPR009729
Galactosyl transferase	IPR002659
Glycosyl hydrolase family 59*	IPR001286
GoLoco motif	IPR003109
Heme NO binding associated	IPR011645
Heparan sulfate 2-O-sulfotransferase*	IPR007734
Hormone receptor	IPR000536
Hormone-sensitive lipase (HSL) N-terminus	IPR010468
HYR*	IPR003410
Immunoglobulin	IPR013151
Immunoglobulin c-2*	IPR003598
Immunoglobulin I-set*	IPR013098
Immunoglobulin V-set*	IPR013106

Integrin alpha	IPR013519
Inward rectifier potassium channel	IPR013521
Kunitz/bovine pancreatic trypsin inhibitor*	IPR002223
L27	IPR004172
Laminin G*	IPR001791
Laminin N-terminal	IPR008211
Leucine rich repeat N-terminal	IPR000372
Low-density lipoprotein receptor class A	IPR002172
Mbt repeat	IPR004092
MOFRL family*	IPR007835
N-AcetylglucosaminyltransferaseIV(GnT-IV) conserved region	IPR006759
Nebulin repeat	IPR013998
Nine cysteines of family 3 GPCR	IPR011500
NRF (N-terminal domain in C. elegans NRF-6)	IPR006621
Organic anion transporter polypeptide (OATP)	IPR004156
p53 DNA-binding	IPR011615
Pacifastin inhibitor (LCMII)	IPR008037
PET	IPR010442
Phosphomevalonate kinase	IPR005919
Phosphotyrosine binding (IRS-1 type)	IPR013625
Phosphotyrosine interaction (PTB/PID)	IPR006020
PHR	IPR012983
PI3-kinase family, p85-binding	IPR003113
Plexin	IPR013548
Progressive ankylosis protein (ANKH)	IPR009887
Protein of unknown function (DUF1241)	IPR009652
Raf-like ras-binding	IPR003116
Reeler	IPR002861
Renin receptor-like protein	IPR012493
Repeat in HS1/cortactin	IPR003134
S-100/ICaBP type calcium binding	IPR013787
Sarcoglycan complex subunit protein	IPR006875
Selenoprotein S (SelS)	IPR009703
Seven transmembrane receptor, secretin family	IPR000832
SH3 domain-binding protein 5 (SH3BP5)	IPR007940
Somatomedin B	IPR001212
Spin/Ssty family	IPR003671
STAT protein, DNA binding	IPR013801
TNF (Tumor Necrosis Factor)	IPR006052
Translocon-associated protein, delta subunit precursor	IPR008855
Tropomyosin	IPR000533
Uncharacterized protein family (UPF0121)	IPR005344
Von willebrand D*	IPR001846
Zinc finger, C2HC type	IPR002515
Fungi and Choanoflagellates	IPR005109
Anp1	IPR005545
YCII-related domain*	IPR005545

*Present in bacteria

Table S6. Species included in comparative protein domain analysis.

Dictyostelium	
<i>Dictyostelium discoideum</i>	<i>Dictyostelium discoideum</i> AX4
Fungi	
<i>Aspergillus fumigatus</i>	<i>Candida glabrata</i>
<i>Cryptococcus neoformans</i>	<i>Encephalitozoon cuniculi</i>
<i>Eremothecium gossypii</i>	<i>Kluyveromyces lactis</i>
<i>Saccharomyces cerevisiae</i>	<i>Schizosaccharomyces pombe</i>
<i>Yarrowia lipolytica</i>	
Metazoa	
<i>Anopheles gambiae</i>	<i>Apis mellifera</i>
<i>Bos Taurus</i>	<i>Caenorhabditis elegans</i>
<i>Canis familiaris</i>	<i>Ciona intestinalis</i>
<i>Danio rerio</i>	<i>Drosophila melanogaster</i>
<i>Gallus gallus</i>	<i>Homo sapiens</i>
<i>Macaca mulatta</i>	<i>Monodelphis domestica</i>
<i>Mus musculus</i>	<i>Pan troglodytes</i>
<i>Rattus norvegicus</i>	<i>Takifugu rubripes</i>
<i>Tetraodon nigroviridis</i>	<i>Xenopus tropicalis</i>
Unicellular eukaryotes	
<i>Cryptosporidium hominis</i>	<i>Cyanidioschyzon merolae</i>
<i>Debaryomyces hansenii</i>	<i>Giardia lamblia</i>
<i>Monosiga brevicollis</i>	<i>Plasmodium falciparum</i>
<i>Thalassiosira pseudonana</i>	

Genomes of these species were used in the initial analysis of the phylogenetic distribution of *M. brevicollis* protein domains. The phylogenetic distributions of domains classified by this analysis as unique to choanoflagellates and another phylogenetic group were manually annotated using the Pfam and SMART online databases.

Table S7. Immunoglobulin domains are restricted to choanoflagellates and metazoans.

	Metazoa			Choanoflagellates	Fungi		Dictyostelia	Plants
	<i>Hsap</i>	<i>Cint</i>	<i>Dmel</i>	<i>Mbre</i>	<i>Ccin</i>	<i>Ncra</i>	<i>Ddis</i>	<i>Atha</i>
	1502	144	503	5	0	0	0	0
Immunoglobulin*								

*Total number of immunoglobulin (Ig)-type domains (Ig, Ig-like, Ig c1-set, Ig subtype 2, Ig v-set) predicted by SMART.

Table S8. Intercellular signaling pathways across phyla.

Pathway	Component	Animals				Choanozoa	Fungi				Amoebozoa	Plant
		<i>Hsap</i>	<i>Cint</i>	<i>Dmel</i>	<i>Nvec</i>	<i>Mbre</i>	<i>Rory</i>	<i>Ncra</i>	<i>Scer</i>	<i>Ccin</i>	<i>Ddis</i>	<i>Atha</i>
NHR	<i>ROR</i>	●	●	●	⊙	○	○	○	○	○	○	○
	<i>Hnf4</i>	●	●	●	⊙	○	○	○	○	○	○	○
	<i>Err</i>	●	●	●	⊙	○	○	○	○	○	○	○
WNT	<i>Wnt</i>	●	●	●	●	○	○	○	○	○	○	○
	<i>Fzd</i>	●	●	●	●	○	○	○	○	○	●	○
	<i>Dsh</i>	●	●	●	●	○	○	○	○	○	○	○
TGFβ	<i>ALK</i>	●	●	●	●	○	○	○	○	○	○	○
	<i>TGFβr</i>	●	●	●	●	○	○	○	○	○	○	○
	<i>Smad</i>	●	●	●	●	○	○	○	○	○	○	○
NFKβ/Toll	<i>NFKβ</i>	●	●	●	●	○	○	○	○	○	○	○
	<i>Tlr</i>	●	⊙	●	●	○	○	○	○	○	○	○
	<i>Tollip</i>	●	●	○	●	●	○	○	○	○	○	○
JAK/STAT	<i>Jak</i>	●	●	●	⊙	○	○	○	○	○	○	○
	<i>Stat</i>	●	●	●	⊙	⊙	○	○	○	○	⊙	○
Notch	<i>Notch</i>	●	●	●	●	⊙	○	○	○	○	○	○
	<i>Delta</i>	●	●	●	⊙	○	○	○	○	○	○	○
	<i>Presenilin</i>	●	●	●	●	●	●	○	○	○	●	●
	<i>Furin</i>	●	●	●	●	⊙	○	○	○	○	○	○
	<i>TACE</i>	●	●	●	●	⊙	○	○	○	○	○	○
Hedgehog	<i>Ptc</i>	●	●	●	●	●	○	⊙	⊙	○	⊙	⊙
	<i>Hh</i>	●	●	●	●	⊙	○	○	○	○	○	○
	<i>Smo</i>	●	●	●	●	○	○	○	○	○	⊙	○
	<i>Fu</i>	●	●	●	●	●	○	○	○	○	●	○
RTK	<i>Rtk</i>	●	●	●	●	●	○	○	○	○	○	○

A filled circle (●) indicates presence of a homolog with strong similarity. A partially filled circle (⊙) indicates a gene with partial similarity (e.g. contains some but not all domains diagnostic of that protein). An open circle (○) indicates no homologs found. ROR, Retinoid-related orphan receptors ; Hnf4, Hepatocyte nuclear factor 4 ; ERR, Estrogen-Related Receptor; Fzd, Frizzled; DSH Disheveled; ALK, Activin-Like Kinase *TGFβr*, *TGFβ receptor*; SMAD, SMA/MAD Mothers Against Decapentaplegic; Tlr, Toll-like receptor; Jak, Janus Kinase; Stat, ; DSL, Delta Serrate Lag-2, Ptc, Patched; Hh, Hedgehog; Smo, Smoothened; Fu, Fused; Sufu, Suppressor of Fused, Rtk, Receptor Tyrosine Kinase.

Table S9. *M. brevicollis* presents a key intermediate in the evolution of MAPK signaling.

		Animal		Choanoflagellate	Fungi		Dictyostelia
Kinase		<i>H.sap</i>	<i>N.vec</i>	<i>M.bre</i>	<i>S.cer</i>	<i>N.cra</i>	<i>D.dis</i>
MAPKKK	MEKK1	•	•	•			
	MEKK2	•	•	•			
	MTK1(MEKK4)	•	•				
	ASK (MEKK5-7)	•	•	•			
	MEKK15	•	•				•
	Mos	•	•				
	Raf	•	•				
	LZK (MEKK12-13)	•	•	•			
	MLK (MEKK9-11)	•	•	•			
	TAO	•	•	•			
	UNCLASSIFIABLE		•	•	•	•	•
MAPKK	MKK1	•	•	•	•	•	•
	MKK5	•	•	•			
	MKK3	•	•				
	MKK4	•	•				
	TOPK	•	•	•			
	UNCLASSIFIABLE			•	•	•	
MAPK	ERK	•	•	•	•	•	•
	ERK5	•	•	•			
	p38	•	•	•	•	•	
	JNK	•	•				
	ERK3	•	•				
	ERK7	•	•	•			•
	NMO	•	•				
	UNCLASSIFIABLE				•	•	

Sequence analysis of the three tiers of kinases from the MAPK module in metazoans (human, sea anemone (*Nvec*; *Nematostella vectensis*), choanoflagellate (*M. brevicollis*), fungi (*S.cer*: *Saccharomyces cerevisiae*; *N.cra*: *Neurospora crassa*) and slime mold (*Dictyostelium discoideum*) shows the emergence of MAPK modules in choanoflagellates and lower metazoans. Kinase subfamilies on the left are from the classification given at kinase.com, based on human kinases.

Table S10. Basal transcription factors present in *M. brevicollis*.

Basal Machinery		<i>H. sap</i>	<i>D. mel</i>	<i>M. bre</i>	<i>S. cer</i>
RNA polymerase II	Rpb1	●	●	●	●
	Rpb2	●	●	●	●
	Rpb3	●	●	●	●
	Rpb4	●	●	⊙	●
	Rpb5	●	●	●	●
	Rpb6	●	●	●	●
	Rpb7	●	●	●	●
	Rpb8	●	●	●	●
	Rpb9	●	●	●	●
	Rpb10	●	●	●	●
	Rpb11	●	●	●	●
	Rpb12	●	●	●	●
	TBP	●	●	●	●
	TBP 2	○	○	●	○
	TFIIA -L	●	●	●	●
	TFIIA -S	●	●	⊙	●
	TFIIB	●	●	●	●
	TFIIE-L	●	●	●	●
	TFIIE-S	●	●	●	●
	TFIIF-L	●	●	○	●
	TFIIF-S	●	●	●	●
TFIIH	XPB	●	●	●	●
	XPB	●	●	●	●
	p62	●	●	○	●
	p52	●	●	○	●
	p44	●	●	●	●
	p34	●	●	○	●
	cdk7	●	●	●	●
	cyclin H	●	●	●	●
	Mat1	●	●	●	●
Co-activators	p8	●	●	○	
Co-activators		·	·		
	PC4	●	●	●	
TFIID	TAF1	●	●	●	●
	TAF2	●	●	●	●
	TAF3	●	●	○	●
	TAF4	●	●	○	●
	TAF5	●	●	●	●
	TAF6	●	●	○	●

	TAF7	●	●	○	●
	TAF8	●	●	○	●
	TAF9	●	●	●	●
	TAF10	●	●	●	●
	TAF11	●	●	○	●
	TAF12	●	●	○	●
Mediator	MED1	●	●	○	●
	MED2	○	○	○	●
	MED3	○	○	○	●
	MED4	●	●	○	●
	MED5	○	○	○	●
	MED6	●	●	●	●
	MED7	●	●	●	●
	MED8	●	●	●	●
	MED9	●	●	○	●
	MED10	●	●	○	●
	MED11	●	●	○	●
	MED12	●	●	○	●
	MED13	●	●	○	●
	MED14	●	●	○	●
	MED15	●	●	○	●
	MED16	●	●	○	●
	MED17	●	●	○	●
	MED18	●	●	○	●
	MED19	●	●	○	●
	MED20	●	●	○	●
	MED21	●	●	●	●
	MED22	●	●	○	●
	MED23	●	●	○	○
	MED24	●	●	○	○
	MED25	●	●	○	○
	MED26	●	●	○	○
	MED27	●	●	○	○
	MED28	●	●	○	○
	MED29	●	●	○	○
	MED30	●	●	○	○
	MED31	●	●	○	●
Chromatin Transactions		·	·	·	·
	CBP(p300)	●	●	⊙	○
	GCN5	●	●	●	●
	ISWI	●	●	●	●
	SWI/SNF	●	●	●	●
	Osa	●	●	⊙	
Elongation factors					
	TFIIS	●	●	●	●
	PAF-1	●	●	●	●

	NELF	●	●	●	
	DSIF	●	●	●	●

Key: ● - present, ⊙ - weak alignment but present, ○ - absent or unidentifiable. Species abbreviations: *H. sap* - *Homo sapiens*, *D. mel* - *Drosophila melanogaster*, *M. bre* - *Monosiga brevicollis*, *S. cer* - *Saccharomyces cerevisiae*.

Table S11: Number of *M. brevicollis* protein models containing transcription factor family specific domains.

Transcription Factor Family	Pfam Domain Id	No. protein models containing domain
BolA-like	PF01722	1
Cold-shock DBD	PF00313	1
HTH	PF01381	1
PC4	PF02229	1
PAH	PF02671	1
STAT DBD	PF02864	1
Tubby-like	PF01167	1
Homeobox	PF00046	2
HSF DBD	PF00447	2
p53 DBD	PF00870	2
RFX DBD	PF02257	2
ZnF NF-X1	PF01422	2
E2F TDP DBD	PF02319	3
MADS/SRF type	PF00319	4
FH	PF00250	8
bZIP	PF07716, PF00170	12
HLH	PF00010	13
Myb DBD	PF00249	14
ZnF CCCH	PF00642	16
ZnF C2H2	PF00096	68
Total:		155

bZip: basic-leucine zipper; DBD: DNA binding domain; E2f-TDP: E2F/DP (dimerization partner) family winged-helix DNA-binding domain; FH: forkhead; Hbx: homeobox; HLH: helix-loop-helix; HSF: heat shock factor; HTH: helix-turn-helix; PAH: paired amphipathic helix; RFX: regulatory factor X; SRF: serum response factor; STAT: signal transducer and activator of transcription; ZnF: Zinc finger.

Supplementary Notes

S1. Genome sequencing and assembly

S1.1 Pilot sequencing efforts. The bacterivorous lifestyle of choanoflagellates and the lack of robust axenic cultures presented a challenge for the production of a high quality genome sequence and assembly. Pilot sequencing from total genomic DNA preparations (containing both bacterial and *M. brevicollis* DNA) revealed that over 80% of the DNA was bacterial, meaning that coverage of the choanoflagellate genome would be insufficient for a quality assembly. We therefore employed two strategies for dealing with bacterial contamination prior to sequencing: (1) reduction of bacterial diversity in cultures and (2) separation of bacterial and choanoflagellate DNA after DNA isolation. Using physical separation techniques combined with antibiotic treatments, a culture line with only a single contaminating bacterial species, *Flavobacterium* sp., was developed. The GC content of *Flavobacterium* (33%) is sufficiently different from that of *M. brevicollis* (55%) to allow separation of the two genomes over a CsCl gradient. *M. brevicollis* genomic DNA isolated in this manner was used to construct replicate libraries containing inserts of 2-3 kb, 6-8 kb, and 35-40 kb, each of which was used for paired end shotgun sequencing. The estimated fractions of bacterial clones in the main libraries (BHUH, BIFH, BNUS) ranged from 3% - 12% and sequences from these clones assembled almost entirely into a single 4.2 Mb scaffold, presumably representing the full genome of *Flavobacterium* sp.

S1.2 Generation of a monoxenic *M. brevicollis* culture, MX1. *M. brevicollis* (ATCC 50154) grown with mixed bacteria was propagated at 25°C in ATCC 1525, growth media prepared by infusing seawater with Ward's Cereal Grass Media (Ward's Natural Science) until the culture reached stationary growth (four days). To reduce the bacterial diversity, the culture was treated with 50ug/mL streptomycin, 50 ug/mL kanamycin, and 12.5 ug/mL chloramphenicol, supplemented with γ -irradiated *Enterobacter aerogenes*, and then cultured in the dark with gentle shaking for 48 hours. The culture was split and the antibiotic treatment was repeated four additional times. The antibiotic-treated culture was pelleted at 4K rpm, 20 min, 15°C and cultured for 48 hours in antibiotic free ATCC 1525 media, during which there was no apparent bacterial proliferation. Cells from an isolated colony of *Flavobacterium* sp. were then added to the culture to support choanoflagellate growth. The culture was further sterilized via a U-tube technique of migration-dilution adapted from Claff, 1940⁵. Briefly, 15mL of culture were concentrated by centrifugation at 6k rpm for 10 min at 25°, and then resuspended in 5mL of ATCC 1525 media. The concentrated culture was placed in the first well of a six well plate, which was connected by three sterile glass U-shaped tubes to the adjacent well filled with fresh ATCC 1525 media. After 48 hours, the culture in the second well was supplemented with cells from a colony of *Flavobacterium* sp. The resulting culture, MX1, was shown to be monoxenic by PCR amplification, cloning and sequencing of multiple independent bacterial 16S rRNA clones using the following primer set: 5'- AGA GTT TGA TCC TGG CTC AG-3' and 5'-ACC TTG TTA CGR CTT-3', modified from Weisburg et. al, 1991⁶. All clones were identical and related to 16S sequences from bacteria in genus *Flavobacterium*. Members of this genus have GC contents ranging from 31.6%-50.0%⁷.

S1.3 Isolation of *M. brevicollis* genomic DNA. *M. brevicollis* MX1 was grown to a density of 10^7 cells/mL in ATCC 1525 media and 750mL of culture was pelleted by two rounds of centrifugation at 10K rpm for 30 min at 4°C. Cell pellets were frozen at -80°C and ground to a fine powder under liquid N₂. *M. brevicollis* genomic DNA (at this point contaminated with *Flavobacterium sp.* genomic DNA) was isolated with the Puregene® DNA purification system (Gentra Systems). The *M. brevicollis* genomic DNA was separated from the contaminating *Flavobacterium sp.* DNA via CsCl density gradient ultracentrifugation. Briefly, 2280ug of contaminated genomic DNA was centrifuged to equilibrium (65K rpm for 40hrs) on six gradients of 1.69g/mL CsCl, in the presence of 40ug/mL of the dye Hoechst 33258 (Molecular Probes). The lower of two resulting bands in each gradient was recovered and the DNA was separated from the Hoechst dye by five extractions with NaCl-saturated n-butanol. The CsCl was dialyzed out of the DNA solution through Spectra/Por® MWCO 8000 dialysis tubing (Spectrum Laboratories, Inc.) over 50 hours at 4°C. The purified *M. brevicollis* genomic DNA was rescued from the dialysis tubing and then ethanol precipitated using Pellet Paint® Co-precipitant (Novagen). The final yield was 24ug of purified *M. brevicollis* genomic DNA, representing a 1% recovery from the initial amount of contaminated genomic DNA. This process was repeated to obtain a sufficient amount of choanoflagellate genomic DNA to build the DNA libraries necessary for sequencing.

S1.4 Genome assembly and validation. The initial data set was derived from 6 whole-genome shotgun (WGS) libraries: two with theoretical insert sizes of 2-3 KB, two with theoretical insert sizes of 6-8 KB, and two with theoretical insert sizes of 35-40 KB (Table S1). The reads were screened for vector using Cross_match (<http://www.phrap.org/phredphrap/phrap.html>), then trimmed for vector and quality⁸. Reads shorter than 100 bases after trimming were excluded.

The data was assembled using release 2.9.2 of Jazz, a WGS assembler developed at the JGI⁸⁻¹⁰. A word size of 13 was used for seeding alignments between reads. The unhashability threshold was set to 40, preventing 13-mers present in the data set in more than 40 copies from being used to seed alignments. A mismatch penalty of -30.0 was used, which will tend to assemble together sequences that are more than about 97% identical. The genome size and sequence depth were initially estimated to be 50 MB and 8.0, respectively.

S1.5 Assembly analysis and quality control. The initial assembly contained 47.4 MB of scaffold sequence, of which 3.7 MB (7.8%) was gaps. There were a total of 1,151 scaffolds, with a scaffold N/L50 of 13/1.10 MB, and a contig N/L50 of 220/52.4 KB. (N50 is the number of pieces (scaffolds or contigs) that account for 50% of the assembly; L50 is the minimum length of these pieces). The assembly was then filtered to remove short and redundant scaffolds:

- Short scaffolds were defined as those with < 1 KB total length.
- Redundant scaffolds were defined as those with < 5 KB total length, where > 80% matched a scaffold that was > 5 KB total length in a single, BLAT-determined alignment (Kent 2002), at any % ID.

After excluding redundant and short scaffolds, there remained 46.0 MB of scaffold sequence, of which 3.4 MB (7.4%) was gaps. The filtered assembly contained 232 scaffolds, with a scaffold N/L50 of 12/1.13 MB, and a contig N/L50 of 210/53.3 KB. The sequence depth derived from the assembly was 8.45 ± 0.09 .

There were 107,459 reads that were not placed in the assembly for various reasons, 13,215 of which were excluded due to quality/vector trimming. Of the remaining 94,244 unplaced reads, the overwhelming majority (~95%) had GC contents that suggested they were part of the *M.brevicollis* genome. The unplaced reads whose mean GC contents were greater than 40% contained roughly 14 MB of trimmed sequence. If this sequence were at the same depth as the rest of genome, it would correspond to roughly 1.7 MB of genome, and so could account for at most about half of the gap sequence. The remainder of the gaps could consist of uncloned segments of the genome, the short/redundant scaffolds, mis-estimates of the gap sizes, or other mis-assembly-related issues.

To estimate the completeness of the original assembly (i.e. including short and redundant scaffolds), a set of 29,246 *M. brevicollis* ESTs was BLAT-aligned to the unassembled trimmed data set, as well as the original assembly itself¹¹. 28,821 ESTs (98.5%) were more than 80% covered by the raw sequence data, 29,053 (99.3%) were more than 50% covered, and 29,139 (99.6%) were more than 20% covered. By way of comparison, of the 29,019 ESTs (99.2%) that had BLAT alignments to the original assembly, 28,387 (97.1%) were more than 80% covered by scaffold alignments, 28,866 (98.7%) were more than 50%, and 28,987 (99.1%) were more than 20% covered.

The mitochondrial genome was available before the assembly was run¹² and was used to identify the corresponding organelle scaffolds. There were three such scaffolds (scaffold IDs 243, 254, and 558) in the released assembly. These scaffolds were excluded from the subsequent genome annotation.

To identify additional contaminant scaffolds, a “kitchen-sink” megablast against the NCBI nt database was performed (using the following parameters: -D 2 -z 1e9 -F "m D" -b 100 -v 100 -p 90 -e 1e-10). The resulting alignments were partitioned by top-level NCBI taxonomic classification: Archaea, Bacteria, Eukaryota, Viroids, Viruses, Other, and Unclassified. The last four were grouped together as “Non-Cellular”, while Archaea and Bacteria were lumped together as “Prokaryotic”. Each scaffold was then tentatively classified based on the distribution of its hits between these three larger categories. Scaffolds with only Eukaryota hits, or no alignments at all, were assumed to be part of the main genome. Scaffolds with some (or all) of their alignments in the other categories had those hits manually examined to determine how reliable they were likely to be. Low-quality hits, or ones to sequences that were probably mislabeled in NCBI, were discounted, and the scaffolds were reclassified based on the remaining ones.

Six scaffolds had various types of non-cellular alignments. Examination of these alignments revealed that four of these scaffolds were almost certainly part of the main genome, due to the nature of the hits themselves, and extensive additional alignments to *M.brevicollis* ESTs. One of the scaffolds (scaffold ID 58) was confirmed as non-cellular material, as it was entirely covered by high % ID alignments to various types of cloning vector. The final scaffold in this set (scaffold ID 170) was tagged as mis-assembled, as it was a chimera of sequences that aligned (on one side) to cloning vectors and *E.coli*, and on the other to eukaryotic sequences. The non-cellular and mis-assembled scaffolds were excluded from the subsequent genome annotation.

Five scaffolds had a combination of eukaryotic and prokaryotic BLAST hits. Examination of the details of these alignments, along with hits to the *M.brevicollis* ESTs, indicated that four of the five (scaffold IDs 16, 31, 43, and 49) were probably part of the

main genome. The fifth (scaffold ID 243) was separately determined to be part of the mitochondrion; see above for details.

Two scaffolds had only prokaryotic hits to the NCBI nt database. Examination of the alignments, and the fact that their GC contents were consistent with the known low-GC prokaryotic contaminant, indicated that they were true prokaryotic scaffolds. One of these scaffolds (scaffold ID 1) was 4.2 MB in length and, as mentioned above, likely represents almost the entire genome of the prokaryotic contaminant.

Finally, seven additional scaffolds (scaffold IDs 56, 62, 99, 171, 221, 233, and 460), while not having any BLAST hits to the NCBI nt database, had GC contents consistent with the known prokaryotic contaminant. Five of these scaffolds (62, 99, 171, 221, and 460) had no BLAT alignments to the *M. brevicollis* ESTs, and so were immediately moved into the prokaryotic contaminant category. The other two scaffolds had some EST alignments (scaffold 56: 75 EST alignments; scaffold 233: 9 EST alignments). However, as even the largest confirmed prokaryotic scaffold had seven EST alignments, the remaining two low-GC scaffolds were moved into the prokaryotic category as well. All of the prokaryotic scaffolds were excluded from the subsequent genome annotation. After the removal of these and the other scaffolds mentioned above, 218 putative nuclear scaffolds remained.

S1.6 No detectable single nucleotide polymorphism in *M. brevicollis*. To characterize the level of variation in the population isolate of *M. brevicollis* that was used for sequencing, we searched for single nucleotide polymorphisms (SNPs) among the whole-genome shotgun (WGS) and expressed sequence tag (EST) reads generated by the sequencing project. Raw sequencing reads were trimmed for vector and quality as described above (S1.4 Genome assembly and validation), leaving 551,090 WGS reads and 29,246 reads available for comparison. To determine the overlapping positions that could be used for SNP detection, we aligned trimmed reads against the JGI *M. brevicollis* genome assembly v1.0 using BLAT v. 32¹¹ with default parameters. A total of 495,647 WGS reads and 28,997 EST reads were successfully mapped to genomic scaffolds. We applied two filters to eliminate incorrect read alignments. First, to ensure unique alignments, we only accepted the best alignment for a read if the ratio between the BLAT score of the second highest scoring alignment and the BLAT score of the highest scoring alignment was no greater than 0.8. Second, we required that paired end reads from the same insert align on the opposite strand to the same genomic scaffold, and within the insert size of the library from which the reads were sequenced. After this filtering step, 388,890 WGS reads and 20,934 EST reads remained for SNP detection.

To produce tractable sets of reads for multiple sequence alignment, we divided the genome into 5 kilobase segments, and produced alignments for each segment using all passing reads either partially or fully included in the segment. Repetitive regions of the genome that have been incorrectly collapsed by the assembly process would cause spurious SNPs to be detected, as reads from two different regions of the genome would be included and aligned within the same segment. To eliminate such segments from consideration, we counted the number of reads mapped by BLAT within each segment with greater than 300 matches to the segment, including all alignments from all trimmed reads, as the uniqueness criterion may have eliminated reads from potentially repetitive regions. More than 90% of segments contained between 0 and 100 reads, and we rejected segments containing 100 or more reads (the average number of reads in a rejected

segment was 747). We created multiple sequence alignments for passing segments using MAP¹³, with a match score of 1, a mismatch score of -2, a gap open cost of 4, a gap extension cost of 3, and a gap limit of 5. To remove alignment artifacts caused by simple repetitive sequence, we did not consider bases within regions detected by Tandem Repeats Finder version 4.00¹⁴, run with the default parameters. We eliminated low quality regions within reads by applying the quality criteria of the Neighbourhood Quality Standard^{15, 16}. Any positions with at least two different alleles passing NQS(25, 20) were considered to be putative SNPs. Using our technique, it is also possible to discover insertions or deletions among WGS and EST reads. However, such differences are significantly more likely to be artifacts of alignment or incorrect base calling, and so we chose to focus our initial variation discovery efforts on SNPs.

We discovered 6,313 putative SNPs among the combined WGS and EST reads, or roughly one SNP per 6,595 sequenced bases. However, the distribution of putative SNP positions in the genome was highly non-uniform, with 4,585 of the putative SNPs within 100 bases of each other. While it is possible that this distribution of SNPs is caused by inhomogeneity in mutation rate or exists due to the action of positive or negative selection, the simplest explanation is that the SNPs within 100 bases of each other are artifacts of over-collapsed regions within the genome assembly that were able to escape our filtering process. Manual examination of 20 randomly selected segments containing two or more SNPs within 100 bases of each other confirmed that all such segments were the result of comparison between two different genomic regions. After eliminating such segments from consideration, only 1,478 putative SNPs remained. In addition, none of these putative SNP positions had more than one read carrying the alternate allele, implying either that all putative SNPs were artifacts of the cloning and sequencing process or that they were present at very low allele frequencies. Manual examination of 20 randomly selected SNPs from the remaining 1,478 putative SNPs revealed 9 of the SNPs to be errors made by the base caller. To investigate the remaining 11 randomly selected SNPs that were not base calling errors, we designed PCR amplicons of roughly 650 bases in length flanking each of the SNPs, and performed PCR followed by sequencing for each amplicon in 4 separate populations of *M. brevicollis*. None of the putative SNP positions was polymorphic in any of the sequenced populations, and no detectable variation was present at any other position within the amplified regions. Thus, our results are consistent with a lack of single nucleotide polymorphism in the sequenced isolate of *M. brevicollis*, although it is formally possible that there is extremely rare variation that our methodology was unable to detect.

S1.7 Mode of reproduction and ploidy of *M. brevicollis* remain unknown. We could not use the lack of variation detected in *Monosiga* to infer ploidy or to determine mode of reproduction. Two strong population bottlenecks occurred in the demographic history of the sequenced culture: one at the initial isolation of *Monosiga* and another during the preparation of a monoxenic strain for sequencing (Supp. Notes S1.2). These bottlenecks may have reduced the population size to two or fewer individuals, and were sufficient to obscure any signal in variation that could have been used to make inferences regarding ploidy or sex. Although our lab cultures were rapidly expanded following both bottlenecks, they retained a small effective population size¹⁷. Therefore, genetic drift could have quickly eliminated variation completely in either a haploid or a diploid

population, given that the relative difference in rate of reduction of heterozygosity is only two-fold¹⁸.

S2. Joint Genome Institute (JGI) annotation of the genome. The JGI annotation pipeline takes multiple inputs (scaffolds, repeats, and ESTs) and produces annotated gene models and other features that are deposited in a database. The data can be accessed by the public through the JGI *M. brevicollis* genome portal at <http://www.jgi.doe.gov/Mbrevicollis>.

Before gene prediction, the 218 scaffolds were masked using RepeatMasker (<http://www.repeatmasker.org/>) and a custom repeat library of 108 putative transposable elements, which are available on the *M. brevicollis* genome portal downloads page. After masking, a variety of gene prediction programs were deployed, based on a variety of methods. These were 1) the *ab initio* method FGENESH¹⁹ (Softberry Inc., NY, USA), the homology-based methods FGENESH+¹⁹ (Softberry Inc., NY, USA) and GeneWise²⁰ seeded by BLASTx alignments against sequences of all opisthokont entries in the GenBank nonredundant protein database as of May 2006, and 3) mappings of EST cluster consensus sequences from *M. brevicollis* produced using EST_map (Softberry Inc., NY, USA). EST clusters were assembled using single link clustering at 98% identity. Both the JGI ESTs and ESTs from ChoanoBase (<http://mcb.berkeley.edu/labs/king/blast/>) were used to assemble clusters.

GeneWise models were completed by using scaffold data to find in frame upstream start and downstream stop codons. EST clusters were used to extend, verify, and complete the predicted gene models using custom scripts (estExt, I. Grigoriev, unpublished). The resulting set of models was then filtered for the “best” models, based on criteria of completeness, length, EST support, and homology support, to produce a non-redundant representative set. This representative set was subject to community-wide manual curation and comparative genomics studies.

9196 non-redundant gene predictions constitute release 1.0. The majority of these genes (87%) were predicted by the *ab initio* method FGENESH using a parameterization based on *M. brevicollis* full-length mRNAs and EST cluster consensus sequences that appeared to contain a full open reading frame. Only 13% of gene structure models were predicted using homology-based methods, specifically FGENESH+ and GeneWise using peptides from GenBank to seed the non-redundant database (Supp. Table S1). When possible, these predictions were corrected and/or extended using ESTs. A small number of gene models (< 1%) were predicted based only on clusters of overlapping ESTs that consistently aligned to the genome and had substantial open reading frames. Though many genes were predicted by *ab initio* methods, the gene catalog is supported by other evidence (Supp. Table S2). 90% of the predicted genes are complete models in the sense of having start and stop codons, 83% of the gene catalog aligns with proteins in the GenBank nr database (e-value < 0.1) and 56% of the predicted genes possess Pfam domains. Furthermore, 46% of the gene catalog is consistent with the ESTs collected from exponentially growing *M. brevicollis*.

All predicted gene models were annotated for protein function using domain prediction tool InterProScan²¹ and hardware-accelerated double-affine Smith-Waterman alignments (<http://www.timelogic.com>) against Swiss-Prot²², KEGG²³, KOG²⁴. Then KEGG hits were used to map EC numbers, and EC, Interpro, and Swiss-Prot hits were

used to map Gene Ontology (GO) terms²⁵. In addition we ran SignalP²⁶ and TMHMM²⁷ for analysis of protein localization.

We predicted that 2,030 proteins (22%) possess a leader peptide, 2,100 proteins (23%) possess at least one transmembrane domain, and 1,132 (12%) possess both. We assigned 1,843 distinct GO terms to 4,834 proteins (53%) using EC-to-GO, Swiss-Prot-to-GO, and InterPro-to-GO mappings (<http://www.geneontology.org/GO.indices.shtml>). We also assigned 1,952 proteins (21%) to KEGG pathways, with a total of 640 distinct EC numbers. The top 4 most populated KEGG pathways are amino acid, complex carbohydrate, carbohydrate, and complex lipid metabolism (436, 387, 289, and 377 proteins, respectively). The complex carbohydrate metabolism pathway includes nearly 200 proteins devoted to the KEGG map starch and sucrose metabolism (MAP00500). Finally, we assigned 6883 proteins (75%) to 3389 KOGs.

S3. Analysis with an evolutionary perspective

S3.1 Phylogenetic Analysis. A previously published 32-species, 50-gene data matrix²⁸ containing metazoan, choanoflagellate and fungal species was updated with the orthologous genes from the *M. brevicollis* genome. Additionally, the corresponding orthologous genes from a fungus (*Rhizopus oryzae*, phylum Zygomycota), a plant (*Arabidopsis thaliana*), and two protists (*Entamoeba histolytica* and *Dictyostelium discoideum*) were added to increase taxonomic diversity in the data matrix. Orthology was established by the reciprocal best BLAST hit criterion²⁹. Specifically, each gene from each of the additional species was considered a true ortholog if it was the best reciprocal BLAST hit with the corresponding gene in *Homo sapiens*.

All analyses were performed on the amino acid sequences. Genes were aligned with CLUSTALW³⁰. Indels and areas of uncertain alignment were excluded from further analysis. Phylogenies were estimated using maximum likelihood (ML) and maximum parsimony (MP), using PHYML³¹ and PAUP*³², respectively (Supp. Fig. S1). Support was assessed using bootstrap re-sampling with 100 replicates (Supp. Fig. S1). For ML, the model of amino acid evolution utilized was estimated by PROTTEST³³ and enforced in all subsequent analyses. The best-fit model for the 50-gene data matrix was WAG³⁴, with rate heterogeneity among sites (value of the gamma shape parameter alpha = 0.87) and a proportion of sites set to be invariable (value = 0.16). MP analyses were performed with all sites equally weighted and with tree-bisection-reconnection branch swapping. Data matrices and trees are available from the authors on request.

S3.2 Gene structure statistics. *M. brevicollis* gene structure statistics are based on the JGI filtered models gene set. The gene structure statistics for other species were found on their respective genome browser websites: *N. vectensis*: <http://genome.jgi-psf.org/Nemve1/Nemve1.home.html>; *C. intestinalis*: <http://genome.jgi-psf.org/Cioin2/Cioin2.home.html>; *N. crassa*: <http://www.broad.mit.edu/annotation/genome/neurospora/>; *C. cinereus*: http://www.broad.mit.edu/annotation/genome/coprinus_cinereus; *D. discoideum*: <http://dictybase.org>) with the exception of *A. thaliana*, for which gene structure statistics were taken from a comparative genome paper³⁵. Many of the *N. vectensis* gene models in the current release are incomplete, so the statistics given are based on a set of over 1,000 genes whose structures are known from full length mRNA (N. Putnam, personal

communications). The estimated gene number was taken from the *Nematostella vectensis* genome paper³⁶.

S3.3 Intron evolution. To study intron loss and gain in orthologous genes in multiple species, we aligned *M. brevicollis* genes to human (ENSEMBL models release 26.35.1), *Drosophila melanogaster* (BDGP4 ENSEMBL model release 41), *Nematostella vectensis* (JGI v1.0), *Phanerochaete chrysosporium* (JGI v2.0), *Cryptococcus neoformans A* (Broad Institute v3.0), *Arabidopsis thaliana* (TIGR release 5), *Chlamydomonas reinhardtii* (JGI v3.0), and *Tetrahymena thermophila* (TIGR, 2005) genes. In 473 cases, a human gene was found to have a mutual best hit to a gene from each of the other nine species, forming a tentative cluster of orthologous genes to be studied further. We also analyzed introns positions from a subset of these species: *Arabidopsis thaliana*, *Cryptococcus neoformans A*, *M. brevicollis*, *N. vectensis*, and *H. sapiens*. This allowed us to analyze a larger number of intron positions than was possible with the nine species data set. In this subset, 538 human genes had mutual best blast hits to each of the other species. Notably, the average numbers of introns per gene in this set of highly conserved genes was different from the average numbers of introns per gene for the entire genomes (12.4 vs. 7.7 introns/gene in humans, 11.7 vs. 5.8 in *N. vectensis*, 8.8 vs. 6.6 introns/gene in *M. brevicollis*, 6.5 vs. 5.3 in *C. neoformans*, and 8.8 vs. 4.4 in *A. thaliana*).

Gene models are often incomplete in the 5' ends and may have poorly determined splice sites, so we restrict our analysis to regions of highly conserved peptides in the orthologs of all five species. The independent identification of such regions in multiple species provides strong evidence for the accuracy of the gene models in these regions. We built multiple alignments of the orthologous clusters using ClustalW and identified gap-free blocks flanked by fully conserved amino acids. We then identified the annotated splice sites within these regions for all the species, with the additional requirements that 1) none of the peptides have a gap in the alignment closer than 3 amino acids from the splice site and 2) no two different peptides have splice sites at different positions closer than 4 amino acids. Empirically, these requirements are necessary to avoid spurious detection of intron gains and losses due to ambiguities in either the multiple alignment or the gene models' splice sites. Finally, we required that at least 5 amino acids out of 10 in the flanking regions of the splice sites be either fully conserved or have strong functional similarity among all species. In the set of genes from all nine species 1,989 intron splice sites at 1,054 highly reliable positions were identified by these requirements. In the five species set 3,847 intron splice sites at 2,121 conserved positions were identified. Presence or absence of introns at these positions across the two sets of taxa was used to build binary character matrices.

Several methods have been developed to infer the evolutionary history of introns in orthologous genes. To gain a comprehensive view of the possible scenarios of intron evolution in *M. brevicollis* and early metazoans, we used three methods; Dollo parsimony, Roy-Gilbert maximum likelihood, and Csuros maximum likelihood. The results of the Csuros maximum likelihood analysis for the nine species set of introns is shown in Figure 2 in the main text and Supp. Table S5. The results of the other analyses for the nine species set are shown in Supp. Figure S3 and the results for the five species set of introns are shown in Supp. Figure S4. Though the different models infer varying amounts of intron loss and gain for various branches, all three models and both data sets

indicate that the ancestor of choanoflagellates and metazoans was as or more intron rich than *M. brevicollis*. Additionally, all models infer a significant gain of introns between the ancestor of metazoans and choanoflagellates and the eumetazoan ancestor, followed by little if any net intron gain within metazoans.

Dollo parsimony assumes that introns appearing at the same positions in orthologous genes were gained only once and then subsequently lost in as many lineages necessary to fit the observed phylogenetic pattern³⁷. The ancestral state in all cases is a gene without introns. Intron gain and loss events were mapped onto the established species tree using PAUP 4.0b10³².

The Roy-Gilbert maximum likelihood method calculates intron loss rates and incorporates them into the estimation of ancestral intron contents³⁸. This method was applied to the current data set using a PERL implementation written and made available by Jason Stajich and Scott Roy³⁹.

The Csuros maximum likelihood method is a probabilistic model that estimates ancestral intron states and intron gain and loss rates for each branch⁴⁰. This method was applied to the current data set using the Java application intronRates.jar made publicly available by the author (<http://www.iro.umontreal.ca/~csuros/introns/>). This model can also infer a number of “all zero” columns, or introns that were present in an ancestral state but lost in all extant taxa. The results shown here assume that there were no such “all zero” columns, but including “all zero” columns in the model does not dramatically change the results for this data set.

From an analysis of all predicted introns in the *M. brevicollis* genome, we observed that its introns are on average shorter than introns found in metazoans. The distribution of *M. brevicollis* intron lengths shows that there are few extremely long introns (Supp. Fig. S2). To determine how this difference manifests itself in introns found in orthologous positions in *M. brevicollis* and metazoans, we examined 419 introns from the set of orthologous introns described above that are found in *M. brevicollis* and humans (Supp. Fig. S2). The average length of these introns in *M. brevicollis* is 132 base pairs as compared to 3,438 base pairs in humans, and the length distributions are significantly different between the two species (Kolmogorov-Smirnov comparison test, $D = 0.815$, $p < 0.01$).

S3.4 Protein domain content of *M. brevicollis*. The protein domain content of the *M. brevicollis* genome was annotated using Pfam v20^{41, 42} and SMART v5.1⁴³ with standard cutoff values. Two protein sets were annotated, the Monbr1_all_proteins.fasta (with completely identical proteins removed) and the Monbr1_best_proteins.fasta. All the analysis described in the text used the Monbr1_best_proteins.fasta set.

The initial analysis of the phylogenetic distribution of protein domains found in *M. brevicollis* included the species listed in Supp. Table S6. To identify domains found exclusively in choanoflagellates and other phylogenetic groups, lists were generated using the Pfam and SMART annotations of these genomes. The lists of Pfam and SMART domains were combined using Interpro ID numbers to eliminate overlap. The phylogenetic distribution of each domain thought to be unique to *M. brevicollis* and a given phylogenetic group was then checked by hand using the SMART and Pfam databases online in order to include additional species distribution information. The

functions of domains identified as unique to *M. brevicollis* and metazoans were hand-curated.

Many of the domains found exclusively in metazoans and *M. brevicollis* are involved in cell signaling and adhesions in metazoans (Supp. Table S4). For example, Bruton's tyrosine kinase motif⁴⁴, which is involved in the regulation of cell proliferation through tyrosine kinase signaling in metazoans is also found in *M. brevicollis*. The *M. brevicollis* genome contains additional domains involved in tyrosine kinase signaling in metazoans, including the phosphotyrosine binding domain (PTB/PID) and the SH3 domain binding protein 5 domain. The *M. brevicollis* genome also encodes metazoan specific domains associated with the extracellular matrix (ECM). These include the reeler domain (found in the neuronal ECM protein reelin⁴⁵), the ependymin domain (an extracellular glycoprotein found in cerebrospinal fluid⁴⁶), and the somatomedin B domain (found in the blood plasma ECM protein vitronectin⁴⁷). Evidence for these protein domains in choanoflagellates, each of which were previously known only in metazoans, extends their evolutionary history to the last common holozoan ancestor, and raises questions about their ancestral functions.

Over and under-represented protein domains in *M. brevicollis* as compared to humans and *S. pombe* were also identified. This analysis was done using SMART's genomic mode, to avoid over-counting domains due to redundant protein sets. Domains predicted by both SMART and Pfam were included and combined using Interpro ID numbers. The number of times each domain occurred in *M. brevicollis* was compared to its occurrence in *S. pombe* and humans. Significantly different numbers of domains were identified by the Chi-square test and ranked by their p-value. The top 200 significantly over and under represented domains were identified. Two sets of comparisons were made, the first of which counted each domain only once per protein and the second of which counted all occurrences of each domain. The top ten over-represented domains as compared to humans and *S. pombe* are shown in Supp. Fig. S5.

Domains that are over-represented in *M. brevicollis* compared to humans include the FG-GAP domain (Interpro ID IPR013517) and the hyaline repeat, or HYR, domain (Interpro ID IPR003410). The FG-GAP domain, a domain that is found in the extracellular portion of transmembrane proteins (e.g. α -integrins) and that mediates interactions with the ECM⁴⁸, occurs in 35 proteins in the *M. brevicollis* genome and only 24 proteins in the human genome. The hyaline repeat (HYR) occurs in 13 proteins in *M. brevicollis* as compared to only three proteins in humans. This predominantly extracellular domain is found in the human glycoprotein hyaline and the sea urchin protein hyalin, which forms an extracellular scaffold around the developing sea urchin embryo⁴⁹. Notably, the five most significantly over-represented domains in *M. brevicollis* relative to *S. pombe* -- ankyrin (IPR002110), SH2 (IPR000980), tyrosine protein kinase (IPR001245), PDZ (IPR001478) and EGF-like (IPR006209) domains -- are important in numerous metazoan signaling pathways. EGF domains are particularly prominent in metazoan multidomain proteins involved in cell signaling⁵⁰.

The SMART and Pfam annotations of the *M. brevicollis* genome, as well as the complete results of the analysis of over and under represented domains, can be found online at <http://smart.embl.de/Monosigia/index.html>.

S3.5 Analysis of signaling, adhesion and transcription factor families. Text and Interpro domain ID searches using the Joint Genome Institute (JGI) *M. brevicollis* v1.0 genome browser (<http://shake.jgi-psf.org/Monbr1/Monbr1.home.html>) were performed to examine the predicted protein models for annotations in categories related to adhesion, signaling, and transcriptional regulation. The online Pfam and SMART tools were used to confirm the presence of domains present in their respective databases. A model was said to contain the domain if both tools identified that domain, except in cases where the domain was not in either the SMART or Pfam database. In these cases, presence predicted by either SMART or Pfam was considered sufficient.

tBLASTn was used to search for members of the transcription factor families listed in Figure 3. All hits with an e-value less than 1 were examined by a reciprocal BLAST search against the NCBI nr (non-redundant) protein database. Those protein models that had reciprocal BLAST hits belonging to the specific transcription factor family were further examined by the Pfam and SMART queries described above if family specific DNA-binding domains were available. Some protein models were further examined if Pfam and SMART did not contain domains specific to the DNA binding domains of the families. The categorization of MbMyc was confirmed by a reciprocal BLAST search against the NCBI nr protein database in which the best defined hits (e.g. not “hypothetical protein”) were all to Myc transcription factors. The *M. brevicollis* Sox transcription factor, found in a tBLASTn search using animal Sox protein sequences, was confirmed by a reciprocal BLAST search against the NCBI nr protein database in which the best defined hits were all to Sox transcription factors.

The presence of specific proteins or domains in *H. sapiens* and *D. melanogaster* was determined by text search in Homologene and Entrez (NCBI). Domains were identified in *C. intestinalis* and *N. vectensis* using the JGI *Nematostella vectensis* v1.0 and *Ciona intestinalis* v2.0 genome browsers (*N. vectensis*: <http://genome.jgi-psf.org/Nemve1/Nemve1.home.html>; *C. intestinalis*: <http://genome.jgi-psf.org/Cioin2/Cioin2.home.html>). Specific proteins and domains in *S. cerevisiae* and *D. discoideum* were identified by text search and GO on their respective genome browsers (<http://www.yeastgenome.org> and <http://dictybase.org>). Specific proteins and domains in the *R. oryzae*, *N. crassa*, and *C. cinereus* genomes were identified by text and BLAST searches of the Broad Institute’s genome browsers (*R. oryzae*: http://www.broad.mit.edu/annotation/genome/rhizopus_oryzae/Home.html, *N. crassa*: <http://www.broad.mit.edu/annotation/genome/neurospora/Home.html>, *C. cinereus*: http://www.broad.mit.edu/annotation/genome/coprinus_cinereus/Home.html). Domains in the *A. Thaliana* genome were identified by BLASTp searches performed on the *Arabidopsis thaliana* Integrated Database (<http://atfdb.org/cgi-perl/gbrowse/atibrowse>).

S3.6 Protein identification numbers for *M. brevicollis* and metazoan signalling homologs. The following *M brevicollis* protein models were identified as homologs of metazoan signaling proteins (JGI protein identification numbers): *Mbrev Tollip*: 38093; *Mbrev STAT-like*: 44371; *Mbrev Notch-like*: 26647; *Mbrev Presenilin*: 29512; *Mbrev Furin-like*: 14515; *Mbrev TACE-like*: 22277; *Mbrev Patched*: 38011, 36995, 36866; *Mbrev Hedgehog-like*: 33852, 36484, 28599; *Mbrev Fused*: 29411.

For the study of Notch and Hedgehog evolution, the following *M. brevicollis* protein models were used: (JGI protein identification numbers): *Mbrev* N1 29255; *Mbrev*

N2 26647, *Mbrev* N3 27644, *Mbrev* H1 28599, *Mbrev* H2 33852. The following metazoan protein sequence were used: (NCBI accession numbers): *Nvec* Notch 20239, *Nvec* Hh 241466, *Nvec* *Hedgling* 200640, *Hsap* Notch NP_060087.2, *Hsap* Hh NP_00184.1

S3.7 Phospho-tyrosine signaling machinery. We used the SMART domain prediction algorithm to assign domain architectures to the proteins in the *M. brevicollis* filtered gene set (filtered SMART set). Within this set we identified all pairwise domain combinations, i.e. the set of domains that appear in the same protein as a TyrKc domain, PTPc domain, or a SH2 domain (Fig. 5). We also performed the pairwise domain analysis for metazoans and non-metazoans (fungi, amoebae, etc.) using the SMART genomic database. Along with the pairwise domain analysis we sorted the filtered set, the metazoan set and the non-metazoan set based on domain architecture of complete proteins using the SMART domain architecture inquiry tool.

S3.8 TATA-binding proteins and transcription elongation factors. *M. brevicollis* possesses a second TATA-binding-protein (TBP) family member, suggesting a choanoflagellate-specific gene duplication that may be associated with gene regulatory diversity. In contrast to the initiation machinery, most of the known eukaryotic transcription elongation factors (TFIIS, NELF, PAF, DSIF, and P-TEFb, but not elongin) have clear homologs in the *M. brevicollis* genome.

S3.9 MAPK signaling. Eukaryotic cells contain multiple mitogen-activated protein kinase (MAPK) cascades that are activated by external stimuli and that produce distinct physiological responses. The core of MAPK signaling is a signature three-kinase module (MAPKKK→MAPKK→MAPK) that is conserved from yeast to human⁵¹. The simple fungal system contrasts with the multiple distinct MAPK pathways in metazoans used to control a larger array of cellular processes. By exploring the MAPK cascade kinases of *M. brevicollis*, we found an unexpectedly early emergence of one MAPK pathway, and potentially new or unstudied variations in the coupling of these pathways.

The canonical Erk MAPK pathway (Mkk1→Erk) is conserved throughout eukaryotes (Supp. Table S9). The functionally distinct Erk5 cascade, (Mekk2→Mkk5→Erk5), was previously found only in deuterostomes⁵², but is now seen in *M. brevicollis*, as well as the primitive metazoan *Nematostella vectensis*, strongly suggesting an ancient origin followed by loss in both insect and nematode lineages⁵³. The evolution of this pathway is intriguing because the three-tiered cascade emerges intact in choanoflagellates with no clear kinase homologs or intermediates in fungi. We do not know the function of Erk5 signaling in choanoflagellates, but in mammals the Erk5 pathway is primarily activated by stress stimuli, and can also be activated by traditional Erk stimuli such as nerve growth factor (NGF)⁵⁴. Erk5 can also be directly activated by PI3 Kinase downstream of the Insulin Receptor.

In contrast with the finding of an intact Erk5 pathway, partial pathway evolution is exemplified by stress-activated p38 MAPK signaling in *M. brevicollis*. A functionally p38-like MAPK is present in yeast (Hog1) and there are at least three clear p38 genes in *M. brevicollis*. These contain the conserved TxY activation phosphorylation site but *M. brevicollis* lacks their canonical activators, Mkk3/Mkk4. This suggests an alternative

upstream kinase of which the best candidate is the dual-specificity kinase TOPK (PBK), which in humans is known to activate p38. This suggests that TOPK might be the original p38 activator and that the Mkk3/Mkk4 kinases evolved more recently within Metazoa. Further evidence for the partial evolution of p38 signaling in choanoflagellates can be found at the MAPKKK level: *M. brevicollis* contains genes not found in fungi encoding apoptosis specific kinase (Ask1), Tao2 and multiple members of the mixed-lineage kinase (MLK) family, kinases that are known to at least partially activate p38 signaling in mammals⁵⁵⁻⁵⁷.

Finally, the choanoflagellate and *Nematostella* genome data reinforce the metazoan-specificity of Jnk signaling. No members of the Jnk MAPK family can be found in fungi or choanoflagellates, and the Jnk activators, Mkk4 and Mkk7, are also missing. Interestingly, many of the MAPKKKs that activate the p38 pathway and the Jnk pathway in mammals are present in *M. brevicollis*. Since Jnk MAPK is most closely related to p38, one hypothesis is that Jnk evolved from a duplication event of p38, and co-opted the upstream components already in place for p38 signaling. Outside of the Jnk pathway, the MAPKs Erk3 and NMO, and the Erk activators Raf and Mos also appear to be exclusive to metazoans.

In summary, MAPK signaling in choanoflagellates is intermediate in complexity between fungi and animals. While *M. brevicollis* lacks some of the hallmarks of metazoan signaling, including p38 activators and the Jnk MAPKs, it has more versatility compared to the fungal MAPK networks, including a full Erk5 cascade and a doubling of the number of MAPKKKs, suggesting a greater diversity of upstream signals and environmental inputs. Future study of the functions of *M. brevicollis* MAPK components will provide an important bridge between the findings from MAPK studies in yeasts and metazoans, and will provide insights into the ancestry and elaboration of the MAPK pathway in animal evolution.

S4. Immunofluorescence Staining of *M. brevicollis*. We fixed *M. brevicollis* cells that were grown shaking at 120 rpm to a density between 10^6 and 10^7 cells/ mL by adding formaldehyde to a final concentration of 4%. We then applied approximately 0.5 mL of the fixed culture to poly-L-lysine coated coverslips and incubated for 30 minutes. After gently washing the coverslips 4 times with PEM (100 mM PIPES, pH 6.9, 1 mM EGTA, 0.1 mM MgSO₄) we blocked and permeabilized the cells for 30 minutes with blocker (PEM/1% BSA/0.3% TritonX-100) and subsequently replaced the blocker with E7 β -tubulin primary antibodies diluted in blocker (Developmental Studies Hybridoma Bank). After incubating the cells with the antibodies for 16 hours at 4° C, we washed the coverslips 4 times with blocker, applied fluorescein conjugated donkey α -mouse IgG (H+L) (Jackson Laboratories) secondary antibodies and incubated for 1 hr in the dark, subsequently washing 4 times with PEM. To visualize F-actin, we incubated the cells with 6 U/ mL rhodamine phalloidin (Molecular Probes) diluted in PEM. To the rhodamine phalloidin-PEM, we added DAPI at a concentration of 10 ng/ mL to visualize the DNA. We applied this mixture to the slides and incubated for 25 minutes in the dark. We then washed the coverslips 3 times with PEM and mounted them onto slides using 10 μ l ProLong Gold antifade reagent (Molecular Probes). All steps were performed at room temperature unless specified otherwise. We took all images using a Leica DMI6000

B inverted compound microscope and Leica DFC350 FX camera at 100X magnification using oil immersion.

S5. Tools for choanoflagellate genomics.

M. brevicollis JGI genome portal:

<http://genome.jgi-psf.org/Monbr1/Monbr1.home.html>

A browser that contains automated and manual gene models and annotations for M. brevicollis. Gene sets and scaffolds can be downloaded.

SMART annotation of *M. brevicollis*:

<http://smart.embl.de/Monosigia/>

SMART protein domain predictions and protein domain architectures for M. brevicollis.

Metazome:

<http://www.metazome.net/>

A multi-taxon tool for comparative genomics.

Choanobase:

<http://mcb.berkeley.edu/labs/king/blast/>

ESTs from the choanoflagellate M. brevicollis and Proterospongia sp.

Taxonomically Broad EST Database:

<http://amoebidia.bcm.umontreal.ca/pepdb/searches/organism.php?orgID=MN>

ESTs from the choanoflagellates Monosiga ovata and M. brevicollis.

References

1. Guillebault, D. et al. A new class of transcription initiation factors, intermediate between TATA box-binding proteins (TBPs) and TBP-like factors (TLFs), is present in the marine unicellular organism, the dinoflagellate *Cryptocodinium cohnii*. *J Biol Chem* 277, 40881-6 (2002).
2. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572-1574 (2003).
3. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754-5 (2001).
4. Csuros, M. in *Proceedings of the Comparative Genomics: RECOMB 2005 International Workshop*; Dublin, Ireland. (ed. McLysaght A, H. D.) 47-60 (Springer-Verlag, Berlin, 2005).
5. Claff, C. L. A migration-dilution apparatus for the sterilization of protozoa. *Physiol. Zool.* 13, 334-341 (1940).
6. Weisburg, W. G., Barnes S.M., Pelletier D.A., and Lane D.J. 16S rDNA amplification for phylogenetic study. *J Bacteriol.* 173, 697-703 (1991).
7. Sottile, M. I., Baldwin, J. N. & Weaver, R. E. Deoxyribonucleic acid hybridization studies on *Flavobacterium meningosepticum*. *Appl Microbiol* 26, 535-9 (1973).
8. Chapman, J. A. in *Physics* (University of California, Berkeley, Berkeley, 2004).
9. Aparicio, S. et al. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* 297, 1301-10 (2002).

10. Putnam, N. H. in *Physics* (University of California, Berkeley, Berkeley, 2004).
11. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* 12, 656-64 (2002).
12. Bullerwell, C. E., Gray, M.W. Evolution of the mitochondrial genome: protist connections to animals, fungi and plants. *Current Opinion in Microbiology* 7, 528--534 (2004).
13. Huang, X. On global sequence alignment. *Comput Appl Biosci* 10, 227-35 (1994).
14. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27, 573-80 (1999).
15. Altshuler, D. et al. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* 407, 513-6 (2000).
16. Mullikin, J. C. et al. An SNP map of human chromosome 22. *Nature* 407, 516-20 (2000).
17. Hartl, D. L. & Clark, A. G. *Principle of Population Genetics* (Sinauer Associates, Inc., Sunderland, MA, 1997).
18. Fisher, R. A. *The Genetical Theory of Natural Selection* (Clarendon, Oxford, 1930).
19. Salamov, A. A. & Solovyev, V. V. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* 10, 516-22 (2000).
20. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Research* 14, 988-995 (2004).
21. Quevillon, E. et al. InterProScan: protein domains identifier. *Nucleic Acids Res* 33, W116-20 (2005).
22. Boeckmann, B. et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 31, 365-70 (2003).
23. Kanehisa M, G. S., Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. From genomics to chemical genomics: new developments in KEGG. *Genome Biology* 5, R7 (2006).
24. Koonin, E. V. et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biology* 5, R7 (2004).
25. Ashburner, M. et al. Gene ontology: tool for the unification of biology. . *Nature Genetics* 25, 25-9 (2000).
26. Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *Journal of Molecular Biology* 340, 783-795 (2004).
27. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology* 305, 567-580 (2001).
28. Rokas, A., Kruger, D. & Carroll, S. B. Animal evolution and the molecular signature of radiations compressed in time. *Science* 310, 1933-8 (2005).
29. Koonin, E. V. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* 39, 309-38 (2005).
30. Thompson, J. D., Higgins, D. G. & Gibson, T. J. Clustal-W - improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22, 4673-4680 (1994).

31. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52, 696-704 (2003).
32. Swofford, D. L. (Sinauer, Sunderland, MA, 2002).
33. Abascal, F., Zardoya, R. & Posada, D. Prottest: selection of best-fit models of protein evolution. *Bioinformatics* 21, 2104-2105 (2005).
34. Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* 18, 691-9 (2001).
35. Town, C. et al. Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy. *Plant Cell* 18, 1348-59 (2006).
36. Putnam, N. H. et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* 317, 86-94 (2007).
37. Kondrashov, F. & Koonin, E. Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from intron sequences. *Trends in Genetics* 19, 115-119 (2003).
38. Roy, S. W., Gilbert, W. Complex early genes. *Proceedings of the National Academy of Sciences* 102, 1986-1991 (2005).
39. Stajich, J. E., Dietrich, F. S. & Roy, S. W. Comparative genomic analysis of fungal genomes reveals intron rich ancestor. (2007).
40. Csuros, M. in *Proceedings of the Comparative Genomics: RECOMB 2005 International Workshop* (ed. McLysaght, A., Huson, D.) 47-60 (Berlin: Springer-Verlag, Dublin, Ireland, 2005).
41. Finn, R. D. et al. Pfam: clans, web tools and services. *Nucleic Acids Res* 34, D247-51 (2006).
42. Bateman, A. et al. The Pfam protein families database. *Nucleic Acids Res* 32, D138-41 (2004).
43. Letunic, I. et al. SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res* 34, D257-60 (2006).
44. Lindvall, J. M. et al. Bruton's tyrosine kinase: cell biology, sequence conservation, mutation spectrum, siRNA modifications, and expression profiling. *Immunol Rev* 203, 200-15 (2005).
45. Tissir, F. & Goffinet, A. M. Reelin and brain development. *Nat Rev Neurosci* 4, 496-505 (2003).
46. Suarez-Castillo, E. C. & Garcia-Ararras, J. E. Molecular evolution of the ependymin protein family: a necessary update. *BMC Evol Biol* 7, 23 (2007).
47. Schwartz, I., Seger, D. & Shaltiel, S. Vitronectin. *Int J Biochem Cell Biol* 31, 539-44 (1999).
48. Baneres, J. L., Roquet, F., Martin, A. & Parello, J. A minimized human integrin $\alpha(5)\beta(1)$ that retains ligand recognition. *J Biol Chem* 275, 5888-903 (2000).
49. Wessel, G. M., Berg, L., Adelson, D. L., Cannon, G. & McClay, D. R. A molecular analysis of hyalin--a substrate for cell adhesion in the hyaline layer of the sea urchin embryo. *Dev Biol* 193, 115-26 (1998).
50. Tordai, H., Nagy, A., Farkas, K., Banyai, L. & Patthy, L. Modules, multidomain proteins and organismic complexity. *Febs J* 272, 5064-78 (2005).

51. Widmann, C., Gibson, S., Jarpe, M. B. & Johnson, G. L. Mitogen-activated protein kinase: conservation of a three-kinase module from yeast to human. *Physiol Rev* 79, 143-80 (1999).
52. Bradham, C. A. et al. The sea urchin kinome: a first look. *Dev Biol* 300, 180-93 (2006).
53. Manning, G., Plowman, G. D., Hunter, T. & Sudarsanam, S. Evolution of protein kinase signaling from yeast to man. *Trends Biochem Sci* 27, 514-20 (2002).
54. Nishimoto, S. & Nishida, E. MAPK signalling: ERK5 versus ERK1/2. *EMBO Rep* 7, 782-6 (2006).
55. Chen, Z. & Cobb, M. H. Regulation of stress-responsive mitogen-activated protein (MAP) kinase pathways by TAO2. *J Biol Chem* 276, 16070-5 (2001).
56. Gallo, K. A. & Johnson, G. L. Mixed-lineage kinase control of JNK and p38 MAPK pathways. *Nat Rev Mol Cell Biol* 3, 663-72 (2002).
57. Matsukawa, J., Matsuzawa, A., Takeda, K. & Ichijo, H. The ASK1-MAP kinase cascades in mammalian stress response. *J Biochem (Tokyo)* 136, 261-5 (2004).